Management of Fraud: Case of an Indian Insurance Company

Sunita Mall[1], Prasun Ghosh[2] & Parita Shah[3]

[1] Assistant professor(Statistics), MICA, Ahmedabad, India

[2] Business Analyst, Hansa Cequity, India

[3] Senior Manager Actuarial, TATA AIG General Insurance Company Ltd., India

Correspondence: Sunita Mall, Assistant professor(Statistics), MICA, Ahmedabad, India

Received: May 12, 2017 Accepted: January 7, 2018 Online Published: April 29, 2018

doi:10.5430/afr.v7n3p18 URL: https://doi.org/10.5430/afr.v7n3p18

Abstract

Frauds in insurance are typically where a fraudster tries to gain undue benefit from the insurance contract by ignorance or wilful manipulation. Using the claims data in motor insurance obtained from a Mumbai based insurance company for the time period of 2010-2016, this study focuses on studying the pattern exhibited by those claims which have been rejected and accepted as well. The prime objective of the study is to identify the important or the significant triggers of fraud and predicting the fraudulent behaviour of the customers using the identified triggers in an existing algorithm. This study makes use of statistical techniques like logistic regression & CHAID (Chi Square Automatic Interaction Detection) technique to identify the significant fraud triggers and to determine the probability of rejection & acceptance of each claim coming in future respectively. Data mining techniques like decision tree and confusion matrix are used on the important parameters to find all possible combinations of these significant variables and the bucket for each combination.

This study finds that variables like Seats/Tonnage, No Claim Bonus, Type of Vehicle, Gross Written Premium, Sum Insured, Discounts, State Similarity and Previous Insurance details are found to be significant at 1% level of significance. The variables like Branch Code and Risk Types are found to be significant at 5% level of signify cance. The Gain chart depicts that our model is a fairly good model. This research would help the insurance company in settling the legitimate claims within less time and less cost and would also help in identifying the fraudulent claims.

Keywords: Fraud, Risk Types, Start Close Proximity, State Similarity, Logistic Regression, CHAID Analysis

1. Introduction

Over the last decade, there is a transformation seen in the Insurance Industry in India. India has become one of the top prospective emerging markets for the insurance sector because of the premeditated view of the largest insurance companies in the world which have found a way in India post Liberalization. The industry has perceived periods of rapid progress along with durations of growth moderation, escalating competition with the segments, life and general insurance having more than 20 competing companies, with substantial growth in customer base. There has been a significant impact of the changing regulatory environment on the development of the industry. After liberalization, the insurance sector has had a steady rise in insurance penetration from 2.71% in 2001 to 5.20% in 2009. A similar movement is observed in the level of insurance density. Looking at the pace at which the insurance sector is growing and also the size of the industry, it is bound to come across a lot of challenges. After a decade of strong growth, the Indian insurance industry is currently facing severe headwinds. The insurance industry seems to be in a state of flux.

Insurance is a policy or contract in which an individual or entity gets the financial protection or reimbursement against the losses from an insurance company. Insurance can be broadly defined as a form of risk management against uncertain loss. The individual or entity gets a payment from the insurer as per the policy details for the loss caused by perils in line with the premium he pays. The company pools client's risks to make payment more affordable for the insured. The insurance policy which the insured gets from the insurer consists of the detail term and conditions of the specific policy and the detailed payment structure.

Life insurance and Non-Life Insurance are two broad categories of insurance industry. Life insurance is a contract between an insured and an insurer, which is based on payment of a sum of money by the insurer against the premium paid by the insured for a specific insurance policy under some specified terms and conditions. Depending on the contract, other events such as terminal illness or critical illness can also trigger payments. The policy holder typically

pays a premium, either regularly or in lump sum. Non-Life Insurance or General Insurance policies provide payments depending on the loss from a particular financial event. General insurance is typically defined as an insurance that is not determined to be life insurance. It is called Property & Casualty Insurance in US & Canada and Non-Life Insurance in Continental Europe.

As the insurance industry in India have been growing in its size and capacity, the challenges for this sector has also become multifold. Pricing, innovation, lapsation, cross-selling, fraud management are some of the striking challenges which is creating a significant hindrance in its growth. According to Ernst & Young survey on frauds in Insurance, the Indian Insurance sector incurs a loss of more than 8% of its total revenue collection in a fiscal year. Further the study indicates the average ticket size of a single fraud ranges between INR 25,000 to INR 75,000. (Fraud in Insurance on rise, survey2010-11) Fraudulent and dishonest claims are a major hazard not only for the insurance industry but also for the entire nation's economy.

The objective of this research paper is to identify the fraud triggers and to predict the fraud behaviour of the customers using these triggers. To predict fraud in the future, claim the triggers are used in an existing algorithm. This paper also focuses on identification of high risk policies at policy inception. For this purpose, Logistics regression model and CHAID analysis is used.

The results of this research paper depicts that seats/Tonnage, no claim bonus, type of vehicle, gross written premium, sum insured, discounts, state similarity and previous insurance details are found to be significant at 1% level of significance. The variables like branch code and risk types are found to be significant at 5% level of significance. It is observed, that the probability of a claim with a particular combination of branch code, third party flag, state similarity, gross written premium and reporting delay is very much likely to be a fraudulent claim whereas the probability of a claim with a specific combination of branch code, type of vehicle, seats/tonnage, branch code and previous insurance details is very much likely to be a non-fraudulent claim. This paper focuses on the node classifiers for a fraudulent claim and a non-fraudulent claim.

In section 2 the literature review is discussed. In section 3 the methodology including data description, sample selection is described. Section 4 depicts the statistical analysis; Section 5 explains conclusion, suggestions and limitations, section 6 shows on the implications of the study.

2. Literature Review

Existing literature on fraud analytics in insurance sector is very thinly documented. Fraud is considered as a second category of white-collar crime in United States declared by The U.S department of Justice, second only to violent cases (Sparrow, 2008). Fraud is relatively an invisible crime and difficult to quantify and detect is argued by Sparrow (2008). He also remarked that it is one of the most serious and important area to be explored for future research. Insurance fraud is costly to individuals and the insuring companies. Insurance companies also lose investment income when a fraudulent claim is filed. Palasinski (2009) remarked that insurance fraud presents financial, societal and humanitarian costs. Out of all types of insurance fraud, the maximum cases are of automobile fraud. It is observed and estimated that 10% to 20% of the automobile insurance claims are fraudulent. The automobile premiums in the United States total $110 billion, this would correlate to an approximate insurance fraud issue of $11 billion annually (Boyer, 2007; Miyazaki, 2009). Research depicted the fraud statistics in an interesting way. It is shown that approximately 49% believe that they would not be caught if they filed a fraudulent claim, 24% of the population believes that it is acceptable to exaggerate the value of an insurance claim, and 11% believe that it is acceptable to submit a claim for damages not actually lost and 30% agree that fraudulent activity will increase during downward economic growth.

The insurers follow many effective strategies to minimize or to remove insurance fraud. However, there is no clarity on the strategic approaches and its effect on the reduction of fraud (Furlan et al., 2011). Research revealed that the insurance fraud is difficult to predict and quantify. Though the insurance frauds are subjective in nature and are of different types, the methods and process adopted by different insurers to study and measure fraud is also not alike (Jay, 2012). Furthermore, fraud is viewed as a relatively invisible crime, making it difficult to accurately assess its impact (Derrig et al., 2006). It is also observed that the current approaches of preventing fraud are broad and ineffective (Palasinski, 2009; Wilson, 2009)

There is neither any clear cut clue available nor any measure documented on which preventive insurance fraud strategies and approaches can be developed. The literature review revealed that all research on fraud analytics are categorized as; numerical strategies, bad faith, behavioural theories, training, patient abuse, economic impact and large-scale fraud. There is still some research gap. The available research is basically in Global context. There is

also little research work available on fraud detection and analytics in Indian context. Mostly the existing literature discusses either the concepts of fraud or the determinants of fraud. However, optimizing fraud, mapping fraud and predicting fraud is a research gap and less researched. The prime objective of this research paper is identifying the triggers of fraud and predicting the fraud using these triggers at the policy inception.

3. Objective

The objective of this research work is formulated in line with the research gap and future scope of the existing researches. The main objectives of this research are as follows;

- Identification of Fraud Triggers
- Predicting fraud using existing algorithm
- Identification of high risk policies at policy inception

4. Methodology

This section explains the data collection methods, the data types, the explanatory variables and the statistical techniques used to answer the objective.

4.1 Data

Claims data of an insurance company is highly confidential. The data analysed in this research paper is the back end claims data of each insurance policy which is written by a Mumbai(India) based insurer. The data contains the details of all the claims which are reported and processed in the company and includes details like occurrence date of the event, reporting date, claim amount, claim review date, surveyor details etc. The data covers the time period 2010 to 2016. The data consists of both fraud and non-fraud claims. The research objective is to find the fraud triggers and to use an existing algorithm along with the fraud triggers to identify the high risk policies at the policy inception. The data initially considered for this research work was 4,54327. The rejected ratio was calculated for various contract types (i.e. Private Vehicle, Commercial Vehicle, etc.) under motor line of business. Rejected ratio is the ratio of the number of rejected claims to the total number of claims. The results find that each contract type shows varied figures and commercial vehicle has the maximum rejected ratio of 28.04%. Also while performing exploratory analysis, it was observed that commercial vehicles had a very high claims ratio. Thus the study focuses only on the commercial motor line of business. Thus 46,175 commercial vehicle motor claims were considered for data analysis in this research paper.

4.2 Explanatory Variables

Referring to the existing literature and in line with the research objective some of the variables like previous insurer details, types of vehicle, diesel_flag, service branch, branch code, paid loss, claim details etc are considered. However, the variables like reporting delay, start close proximity, end close proximity, state similarity, third party flag, and no claim bonus (NCB) difference are not been used in existing studies and thus are our contribution to the existing literature.

Table 1 given below shows explanation of the entire explanatory variable used to detect fraud triggers.

Table 1. Definition of Variables

Sr. No.	Variable	Description
1	MM Code	It is the system coding applied to identify a car on the basis of its make (manufacturing company) and model (specific model).
2	TP Flag	It is a dichotomous variable indicating whether the reported claim is an Own Damage case or Third Party case(0/1)
3	Type of Vehicle	It is the segmentation of cars done on the basis of the purpose of use like ambulance, passenger bus, tractor, etc.
4	Seats/Tonnage	The maximum tonnage of the vehicle. Vehicles with high loading capacity will have tonnage value which is significant and needs to be considered. It is taken in combination with the seating capacity of the vehicle.
5	Risk Types	Basic categorization of goods as to whether its goods carrying or passenger carrying or a part of motor pooling.
6	Branch Code	It denotes the company branch where the policy was underwritten. These branches were regrouped on the basis of the state in which they lie and the claims ratio.
7	Channel Code	The medium of getting the new customer or new business. E.g.: through a broker, through bank etc.
8	Gross Written Premium	The actual amount of premium charged to the policyholder for him to enjoy the benefits in the future.
9	Sum Insured	The maximum amount of money that the insurer is liable to pay the policyholder in case the vehicle is totally damaged.
10	Discount %	It is the additional discount or loading on the premium offered to the policy holder after NCB.
11	No Claim Bonus (NCB)	The discount on premium offered for the next year by company if the policy holder doesn't claim in the present policy period.
12	NCB Difference	It is a binary variable to check whether there is a difference between the actual NCB and expected NCB.
13	Reporting Delay	It's the time lag between date of occurrence of the claim event & date of reporting the claim in the company.
14	Previous Insurer Details	Categorization as to whether the car is new or used and if used then whether it had a previous insurance.
15	Start Close Proximity	It is a binary variable calculated to check whether the claim is reported within the first month of the policy period. (Y/N)
16	End Close Proximity	It is an indicator variable stating whether the claim is reported one month before the end of policy period. (Y/N)
17	State Similarity	It is a binary variable indicating whether the state in which the policy was in written is the same as the state where claim is reported. (Y/N)
18	Age of Vehicle	It is the age of the vehicle which was calculated as the difference between manufacturing year & reporting year of the claim event.

4.3 Statistical Techniques

Fraud is very subjective and is defined differently in different context. In this research paper, all the claims that are rejected only due to fraud are considered as fraudulent claims. The claim decision is considered as a binary event as the claim can be either fraudulent or non-fraudulent, hence can be rejected or accepted. Logistic Regression model is used to identify the significant fraud triggers using specific explanatory variables. Logistic Regression defines the impact of multiple independent variables to predict two dependent variable categories. The dependent variable taken in this research paper is dichotomous in nature and coded as 1 if the claim is fraudulent and 0 if it is not fraudulent.

The logistic regression model is used for analysis

$$\text{Logit } (P)=\ln (\text{odds})=\ln(p/1-p)$$

Logit (P) is the log (to base e) of the likelihood ratio that the dependent variable is 1. Moreover, p can only range from 0 to 1 whereas logit (p) scale ranges from negative infinity to positive infinity. It is also symmetrical around the logit of .5 (which is zero). The logistic regression can be defined as:

$$\text{Logit } [P(X)] = \log [P(X) / 1\text{-}P(X)] = a + b_1 X_1 + b_2X_2 + b_3X_3 + \ldots\ldots + b_nX_n \ldots\ldots\ldots\ldots \text{ Eq. (1)}$$

Where P is the probability that a case is in particular category, bi's are the coefficient to be estimated and X_i's are the explanatory variable. The explanatory variables are MM Code, types of vehicle, tonnage, risk type, branch code, channel code, gross written premium, sum insured, discounts, no claim bonus, etc.

CHAID (**C**hi **S**quare **A**utomatic **I**nteraction **D**etection) technique is a decision tree technique that develops an algorithm which helps in understanding how the predictor variables merge together to determine the outcome of a dependent variable. CHAID analysis is specifically useful for data expressed as categorical variables instead of continuous variable. In CHAID analysis, continuous data can be used, where continuous predictors are split into categories with approximately equal number of observations. In CHAID analysis we can visually see the relationship between the split variables and the associated related factor within the tree.

5. Empirical Findings and Analysis

This section deals with the empirical findings. In section 5.1, we discuss the results of Logistic Regression. The result of CHAID Analysis is explained in section 5.2 and in 5.3 Confusion Matrix of the model constructed is explained.

5.1 Logistic Regression Result

Logistic Regression model results are explained below. A positive coefficient indicates that the probability of a claim to be fraudulent increases/decreases with increasing /decreasing values of corresponding explanatory variables taken for analysis in this research paper. On the contrary, for negative regression coefficient it indicates that increasing/decreasing values of the explanatory variable decreases/increases the probability of a claim to be fraudulent.

For Logistics regression model, Hosmer and Lemeshow (H-L) test is a goodness -of-fit test which divides subjects into ten ordered groups of subjects and compares the number actually in each group (observed) to the number predicted by logistic regression. It is an alternative to chi-square model. The result of H-L test is as follows:

Table 2. Hosmer-Lemeshow Test

Step	Chi Square	Df	Sig
1	15.143	8	0.056

It is clear from Table 2 that the H-L statistic has a significance of 0.056 which implies that there is no statistically significant difference in the observed values and the predicted values using the model. Thus, our model is quite a good fit.

The logistic regression results with coefficient estimate of each explanatory variables are explained in the following table.

Table 3. Significant Variables

SIGNIFICANT VARIABLES	ESTIMATED β COEFFICIENT	p-VALUE	LEVEL OF SIGNIFICANCE
SEATS/TONNAGE	-9.09E-06	0.000186	0.01
NCB.DIFF 1	-5.60E-02	0.052623	0.1
CHANNEL.CODE3	3.67E-01	0.74731	0.1
CHANNEL.CODE8	2.99E-01	0.022749	0.05
CHANNEL.CODE9	-1.09E-01	0.0000496	0.01
CHANNEL.CODE10	8.38E-02	0.056495	0.1
RISK.TYPE3	1.51E-01	0.016921	0.05
RISK.TYPE4	9.37E-02	0.068239	0.1
PREVIOUS.INSURER.DET2	1.71E-01	0.0000011	0.01
PREVIOUS.INSURER.DET 3	3.77E-01	0	0.01
REPORTING.DELAY	-2.61E-04	0.066569	0.1
BRANCH.CODE8	-2.23E+00	0.085107	0.1
BRANCH.CODE19	-1.29E+00	0.036361	0.05
BRANCH.CODE26	-1.15E+00	0.05201	0.1
BRANCH.CODE28	-1.52E+00	0.010223	0.05
BRANCH.CODE40	-1.56E+00	0.008701	0.01
BRANCH.CODE46	-1.13E+00	0.057153	0.1
BRANCH.CODE48	-1.20E+00	0.043235	0.05
STATE.SIMILARITY1	-2.29E-01	0	0.01
TYPE.OF.VEHCILE2	1.80E+00	0.0000003	0.01
TYPE.OF.VEHICLE7	1.41E+00	0.000155	0.01
TYPE.OF.VEHICLE10	1.53E+00	0	0.01
TYPE.OF.VEHICLE20	8.77E-01	0.016997	0.05
TYPE.OF.VEHICLE21	7.11E-01	0.035929	0.05
TYPE.OF.VEHICLE22	1.69E+00	0.0000931	0.01
TYPE.OF.VEHICLE26	1.98E+00	0.0000002	0.01
TYPE.OF.VEHCILE27	4.66E-01	0.043388	0.05
TYPE.OF.VEHICLE28	1.44E+00	0.000631	0.01
TYPE.OF.VEHICLE37	5.00E-01	0.044048	0.05
TYPE.OF.VEHICLE41	4.58E-01	0.096737	0.1
TYPE.OF.VEHICLE42	7.34E-01	0.010975	0.05
TYPE.OF.VEHICLE43	6.60E-01	0.028611	0.05
AGE.OF.VEHICLE	1.25E-02	0.068035	0.1
GWP	-1.76E-05	0.0000021	0.01
SI	1.56E-07	0.007905	0.01
DL	4.26E-03	0.0000177	0.01
MM Code	1.12E-02	0.006201	0.01
TP FLAG	4.23E-04	0.076532	0.1
NCB2	-6.29E-02	0.077093	0.1
NCB3	-1.19E-01	0.007476	0.01
NCB4	-1.01E-01	0.066941	0.1
NCB5	-1.80E-01	0.009327	0.01

The beta (**β**) values are the logistics regression coefficients which is used to create predictive equations. The results in Table 3, show that sixteen variables are found to be significant out of eighteen variables. It is also observed that for a particular variable say Channel Code, four channel codes are found to be significant out of twelve channel codes. Similar inference can be derived for variables having more than one category. The variables like Seats/Tonnage, no claim bonus, type of vehicle, gross written premium, sum insured, discounts, make &model code, state similarity and previous insurance details are found to be significant at 1% level of significance, the variables like branch code and risk types are found to be significant at 5% level of significance whereas channel code, no claim bonus difference, age of the vehicle, third party flag and reporting delay are significant at 10% level of significance. It indicates that the mentioned variables made a significant contribution to predict fraudulent claims at respective level of significance. The variables like start close proximity and end close proximity are not significant predictors. To determine how variables best combine to explain the outcome in a given dependent variable, CHAID analysis is performed and the results is explained in the next section.

5.2 CHAID Analysis Result

CHAID or Chi Square Automatic Interaction Detection is a Classification Tree technique. It tells how variables best merge to explain the outcome in a given dependent variable by building a predictive model. CHAID creates all possible cross tabulations for each categorical predictor until the best outcome is achieved and no further splitting can be performed. In this research paper a nominal CHAID model is fitted on sixteen significant variables as the dependent variable is dichotomous. The results of decision tree obtained after performing CHAID Analysis using sixteen independent variables are as follows:

Table 4. Model Summary

Growing Method	CHAID
Dependent Variable	REJECTED/ACCEPTED
Independent Variable	CHANNEL CODE,RISK TYPE, PREVIOUS INSURER,TP FLAG,REPORTING DELAY, BRANCH CODE, STATE SIMILARITY, TYPE OF VEHICLE, AGE OF VEHICLE, MMCODE,SEATS/TONNAGE, GWP, SI, DL%, NCB,NCB DIFFERENCE
Maximum Tree Depth	10
Minimum Cases in Parent Node	100
Minimum Cases in Child Node	50
Number of Nodes	331
Number of Terminal Nodes	192
Depth	7

It is observed from Table 4, that the maximum number of depth is 7. Total number of terminal nodes is 192 which imply that total number possible combinations of all the independent variables are 192.

CHAID analysis is used to choose the most significant independent variable out of sixteen explanatory variables to bifurcate 42390 heterogeneous data points into homogenous clusters. The criteria used for selection of the independent variable are by using the Chi-Square Statistic. The results are discussed below:

Table 5a. Results of Decision Tree Analysis (Probability of rejection = 1)

High chances of fraud happening				
Before the split		**27.60%**	**72.40%**	
Node Classifier	**Chi-Square Value**	**Percentage of claims rejected**	**Percentage of claims accepted**	**No. of Split Nodes**
Branch Code	1474.199	42.80%	57.20%	9
TP Flag	760.802	87.40%	12.60%	2
State Similarity	487.424	96.60%	3.40%	2
GWP	14.036	97.10%	2.90%	3
Reporting Delay	9.552	99%	1.00%	2

Table 5a exhibits the results of decision tree with probability of rejection equal to 1. Each of the clusters is homogenous within but heterogeneous amongst themselves. The clusters having highest rejection ratios are taken into consideration and further bifurcation is made for the same cluster using next variable. This process continues till all the nodes are the terminal nodes i.e. the tree ends and the data points can't be broken further.

It is observed that the very first variable is branch code with the corresponding chi square value 1474 and the rejection ratio 42.8% which has got 9 splits. The next split occurred for third party flag. Third party flag is a binary variable taking two values i.e. yes or no with a rejection ratio of 87.4% & 33.1% respectively. For further interpretation only YES cluster with highest rejection ratio is considered. Third party flag is broken further using the next variable state similarity with Chi Square = 487.82. State similarity is broken into gross written premium and reporting delay subsequently.

Table 5a also depicts, the probability that a claim coming from the node classifier combination like branch code, third party flag, state similarity, gross written premium and reporting delay is very much likely to be a fraudulent claim whose probability of occurrence is almost equal to 1. Such claims can be directly rejected by the company.

Table 5b. Results of Decision Tree Analysis (Probability of rejection = 0)

Legitimate claim and hence should be accepted				
Before the split		**27.60%**	**72.40%**	
Node Classifier	**Chi-Square Value**	**Percentage of claims rejected**	**Percentage of claims accepted**	**No. of Split Nodes**
Branch Code	1474.199	26.80%	73.20%	9
Type of Vehicle	298.555	39.00%	61.00%	5
Seats/Tonnage	78.579	45.10%	54.90%	2
Branch Code	17.557	43.00%	57.00%	2
Previous Insurer	14.829	34%	66.30%	3

Table 5b displays the results of decision tree with probability of rejection equal to 0. It is observed that the very first variable is with the corresponding chi square value 1474 .19 and the rejection ratio 26.8% which has got 9 splits. The next split is found for Type of Vehicle. Type of Vehicle (Chi-square 298.555) has got five splits and the rejection ratio is 39%. The subsequent node classifiers are Seat/Tonnage (Chi square 78.579and rejection ratio 45.1%), Branch Code (Chi square 17.557and rejection ratio 43%) and Previous Insurer (Chi square 14.829and rejection ratio 34%).

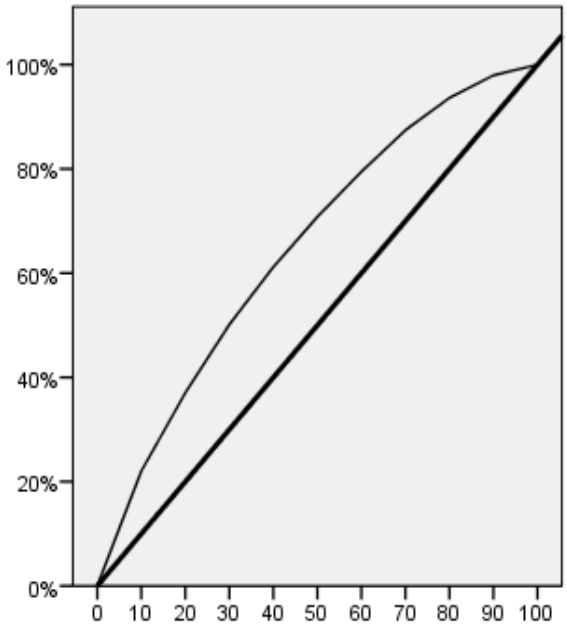
Table 5b also depicts, the probability that a claim coming from the node classifier combination like Branch code, Type of vehicle, Seats/Tonnage, Branch Code and Previous insurer is very much likely to be a non-fraudulent claim.

Goodness-of-fit explains the discrepancy between the observed values and expected values. Any statistical model should be checked with its goodness- of- fit that is how well it fits the data or a set of observations. We checked the best fit of the model by using Gain chart. The result is displayed and discussed below.

5.3 Gain Chart Results

Gain chart is used to check the best fit of the model. Generally, a Cumulative Gain Chart moves from one end to the other with a range of 05 and 100%. For a good model, the gains chart rises sharply toward 100% and then remain stable. On the contrary, a model that provides no information will follow the diagonal reference line. In the Gain chart, X axis represents percentiles and Y axis represents the Gain achieved. The gain chart result is shown in plot 7.

Plot 7. Gain Chart



It is clear from the Gains Chart that our model is a fairly good one. Model performance is also very important which tells how well does the model classify the data points. Confusion matrix is made to understand, evaluate and validate the model performance.

5.4 Confusion Matrix Results

Confusion matrix is used to test the performance of classification model on a data set. It gives an image of the performance of an algorithm. The diagonal elements represent the number of points for which the predicted label is same as the true label whereas the off-diagonal elements are those that are mislabelled by the classification model.

The performance and accurate classification of our model is tested using the confusion matrix. The decision is dichotomous in nature. The decision which is of interest for the study is termed as the positive label and the other as negative label. For our classification model, rejection of claims is considered as positive whereas the acceptance of claim is considered as negative. The threshold value of 0.5 is considered to reject or accept the claim. Any claim with probability greater than 0.5 were rejected and less than 0.5 were accepted by the model. The confusion matrix is as follows:

Table 6. Results of Confusion matrix

0.5		MODEL	
		Rejected	Accepted
COMPANY	Rejected	2201	**10325**
	Accepted	**1203**	31588

For a threshold value of 0.5, our model is found to be 84.56% accurate. The sensitivity at 0.5 cut-off value was 17.57% and specificity was 96.33%. There are 1203 Claims which were accepted by the company but are rejected by the model at 0.5 threshold value. These claims could be accounted for the loss incurred due to settlement of such claims which ideally should have been denied by the company since there was a chance of it being more fraudulent. There are 10325 Claims which were rejected by the company but accepted by the model. There are chances that some loyal

customers with genuine claim might have been denied the claim amount. This would lead to a loss of loyal customers and in turn loss of premium.

6. Conclusion

In this research paper we study the fraudulent behaviour and identified the fraud triggers. We also constructed an algorithm for future claims and identified the high risk policies at policy inception. Little research is done in Global context. This is to the best of our knowledge that there is little research on insurance fraud in Indian context. The variables like Seats/Tonnage, No claim bonus, Type of Vehicle, Gross Written Premium, Sum Insured, Discounts, State Similarity and Previous Insurance Details are found to be significant at 1% level of significance. The variables like Branch Code and Risk Types are found to be significant at 5% level of significance whereas Channel code and Reporting Delay are significant at 10 % level of significance. These variables are significantly contributing to predict fraudulent claims. It is observed, the probability that a claim coming from the node classifiers like Branch code, Third Party Flag, State Similarity, Gross Written Premium and Reporting delay is most likely to be a fraudulent claim whose probability is almost equal to 1 and the probability that a claim coming from the node classifiers like Branch code, Type of vehicle, Seats/Tonnage and Previous insurance details is very much likely to be a non-fraudulent claim. The gain chart depicted that the model built is a fairly good model.

7. Implications for Theory and Practice

This research is very useful for detection and prediction of fraud claims in motor insurance. Similar algorithm can be constructed by the insurer to identify the high risk policies at policy inception. This piece of work will help in reduction of claim processing time. The model developed will get a probability value for each new claim coming in and that will decide the likelihood of the claim to be accepted or rejected. This will reduce the investigation time of claims. This research work will help the insurer to have a consistent approach in handling claims. Reduction in claim reviewing time and manual labor will be reduced and will save out on detailed study of the back papers for each and every claim. This will thus optimize the claim resources and save out on unwanted expense. Impartial decision making and consistent approach will help in retaining loyal customers and the quick processing of claims will reduce the waiting time for the policyholder which will lead to better customer service and provide satisfaction to customers. The combinations that are extracted from the CHAID analysis output of our research paper can provide useful inputs for the underwriting team while writing new business.

The future research scope with this result is very wide. Occurrence of fraud in an insurance company can be broadly categorized as i) internal fraud and ii) external fraud. In this research paper external fraud which is basically committed by the policy holders is highlighted. However, internal fraud which is committed by the employee and agent of the host company is a future scope of research. To identify the important triggers of fraud we have considered the vehicle related, policy holder related and geography related variables. More number of variables can be generated including the employee and agent related characteristics. The fraud is detected and predicted only in the domain of claim. However, the similar research can be done in the domain of premium, underwriting etc. It is beneficial for the insurer to apply a statistical robust model to predict the probability of fraud. This paper has taken a small step in the same direction.

References

Bhoumik Rekha. (2011). Detecting auto insurance fraud by data mining techniques. *Journal of emerging trends in computing and information sciences, 2*(4), ISSN 2079-8407.

Belhadji El Bachir , Dionne Georges. Development of an Expert system for the automatic detection of Automobile Insurance fraud", working paper,ISSN:1206-3304.

Boyer, M. (2007). Resistance (to Fraud) is Futile. *Journal of Risk and Insurance, 74*(2), 461–492. https://doi.org/10.1111/j.1539-6975.2007.00221.x

Furlan, S., O. Vasilecas, & M. Bajec. (2011). Method for Selection of Motor Insurance Fraud Management System Components Based on Business Performance. *Technological and Economic Development Of Economy, 17*(3), 535–561. https://doi.org/10.3846/20294913.2011.602440

Jay, D. (2012). Trend watch: New Developments about Fraud in America. *Journal of Insurance Fraud in America, 3*(2).

J.Holton Wilson. An analytical approach to detecting insurance fraud using Logistics regression. *Journal of finance & Accountancy*.

Palasinski, M. (2009). Testing Assumptions about Naivety in Insurance Fraud. *Psychology, Crime & Law, 15*(6), 547–553. https://doi.org/10.1080/10683160802392444

SABAU Andrei Sorin. (2012). Clustering based Financial Fraud Detection research. *Informatica Economica, 16*(1).

Sithic H.L,Balasubramanian T. (2013). Survey of Insurance Fraud detection using data mining techniques. *International journal of Innovative technology and Exploring Engineering, 2*(3).

Sparrow, M.K. (2008). Fraud in the U.S. Health-Care System: Exposing the Vulnerabilities of Automated Payments Systems. *Social Research, 75*(4), 1151–1180.

Wilson, J.H. (2009). An Analytical Approach to Detecting Insurance Fraud Using Logistic Regression. *Journal of Finance and Accountancy, 1*, 11–15.