American Journal of Mobile Systems, Applications and Services
Vol. 1, No. 3, 2015, pp. 227-232
http://www.aiscience.org/journal/ajmsas

Analytics for Insurance Fraud Detection: An Empirical Study

Carol Anne Hargreaves*, Vidyut Singhania

Business Analytics, Institute of Systems Science, National University of Singapore, Singapore, Singapore

Abstract

Automobile insurance fraud is a global problem. Handling fraud manually has always been costly for insurance companies. Data analytics can play a crucial role in fraud detection and can aid insurance companies to identify fraud. Typically, there are easily more than thirty variables that are used for the fraud analysis. This paper proposes to determine which variables are significant for fraud detection and to provide a framework for the insurance fraud detection. Further, this paper illustrates the business value of data analytics for insurance fraud detection using an empirical study and demonstrates that through a few business rules, the insurance company can accurately identify fraudulent claims which can most likely reduce costs and increase profitability for the company.

Keywords

Fraud Detection, Analytics, Insurance, Significant Variables, Business Rules, Framework

Received: September 9, 2015 / Accepted: September 20, 2015 / Published online: January 20, 2016

1. Introduction

An estimated 10 percent of claims filed with the U.S. insurance industry are fraudulent and typically, only a single digit percentage of the total claims are prevented or recovered as part of claim handling fraud investigation units [8]. What is fraud? The oxford definition of fraud is [2], “Wrongful or criminal deception intended to result in financial or personal gain”. Fraud is increasing exponentially with the easy access of modern technology and communication, resulting in the loss of trillions of dollars worldwide each year.

Automobile insurance fraud is a global problem. Handling fraud manually has always been costly for insurance companies. Data analytics provides an effective way to be more proactive in the fight against fraud and to identify transactions that indicate fraudulent activity or the heightened risk of fraud. Data analytics can play a crucial role in fraud detection and can aid insurance companies to identify fraud. According to Bolton & Hand [1], the appropriate overall strategy for fraud detection is to use a graded system of investigation. Accounts with very high suspicion scores merit immediate and intensive (and expensive) investigation, while those with large but lower scores merit closer (but not expensive) observation.

Techniques for fraud detection are important if we are to identify fraudsters once fraud prevention has failed. In this paper, we will apply statistical hypothesis testing techniques. This paper proposes to determine which variables are significant for fraud detection and to provide a framework for the insurance fraud detection by deriving business rules to help detect fraud. Further, this paper uses empirical auto insurance data and is structured into 6 sections. While Section 1 is the introduction and emphasizes the importance of detecting fraud and identifying business rules that can alert the business when information is extreme and requires further investigation, Section 2 gives a brief literature review, Section 3, the Objective of the Study, Section 4, an overview on the methodology that may be used for fraud detection, Section 5, presents the statistical analysis results such as

* Corresponding author
E-mail address: carol.hargreaves@nus.edu.sg (C. A. Hargreaves)

statistical significance of key factors and recommendations on the derived business rules to be used, to speed up the process of fraud detection, after which Section 6 presents the conclusion.

2. Literature Review

According to Wilson, J.H [3], Automobile insurance fraud is not just a problem in the U.S. but rather it is a global problem as indicated by an analysis of auto insurance fraud in Spain (Artis, [4]). While our focus is on auto insurance, fraud is prevalent with other forms of insurance as well. Insurance companies are realizing the importance of data analytics in the fraud detection space and are hurriedly opting for expensive fraud solutions that are not aligned to the company's weakness and strengths, according to Verma [5]. In fact, Spathis [9], claims that fraudulent financial statements have become increasingly frequent over the last few years. Further, there is an increasing demand for greater transparency, consistency and more information to be incorporated within financial statements. Spathis [10] constructed a model to detect falsified financial statements. He employed a statistical method, with two alternative input vectors containing financial ratios. The reported accuracy rate exceeded 84%. This study substantiates and supports our use of data analytics for fraud detection.

Costons [8], discusses how business rules and anomaly detection are typically the first lines of defense in fraud screening, testing each claim against algorithms that are designed to detect known types of fraud by identifying specific types of patterns. A KPMG Forensics' Fraud Risk Management report states [6], "unlike retrospective analyses, continuous transaction monitoring allows an organization to identify potentially fraudulent transactions on, for example, a daily, weekly, or monthly basis.

Phua [11], studied fraud detection with a focus on how to classify fraudulent claims using skewed data. Results from this study showed good accuracy with success rates as high as 87%. Businesses should therefore use continuous monitoring efforts to focus on narrow bands of transactions or areas that pose particularly strong risks."

Berry [7], also advocates the use of unsupervised learning to find new insights which can improve the supervised learning results. We therefore used statistical hypothesis techniques as our key focus to help detect fraud. Costons [8], concludes that the longer the data analytics techniques are used, there should be a reduction in the frequency of fraud and that the business overall should experience a reduction in average fraud analysis effort per claim handled and a reduction in total payments made, along with a reduction in unallocated loss adjustment expenses incurred to investigate fraud. It is also with this understanding, that this research was initiated.

3. Objective of Study

This paper has three main objectives. Firstly, this paper aims to demonstrate how data analysis can help to detect fraud. Firstly, this paper aims to identify the key significant variables for fraud detection. Secondly, this paper aims to derive business rules from the significant variable information. The derived business rules are required for flagging suspicious records. Then, the business, after being alerted, performs further investigation on these flagged instances and confirms whether they are genuinely fraudulent or not. Thirdly, based on the empirical data used, this paper will present a framework for fraud detection and may be used for similar fraud detection studies.

4. Methodology

4.1. Framework for Fraud Detection

We have developed a framework to assist business users on the process for identifying fraud. Using the data set provided by Angoss Knowledge Seeker software [12], we performed statistical hypothesis testing on 31 variables to identify which variables were significant and could assist in identifying fraud. Once the significant variables are determined, we use these to profile fraudulent and non- fraudulent claims. Further, the significant variables and the fraudulent profile help to derive the business rules to identify future fraudulent claims. The figure 1 below, displays the process taken to identify and detect fraud.

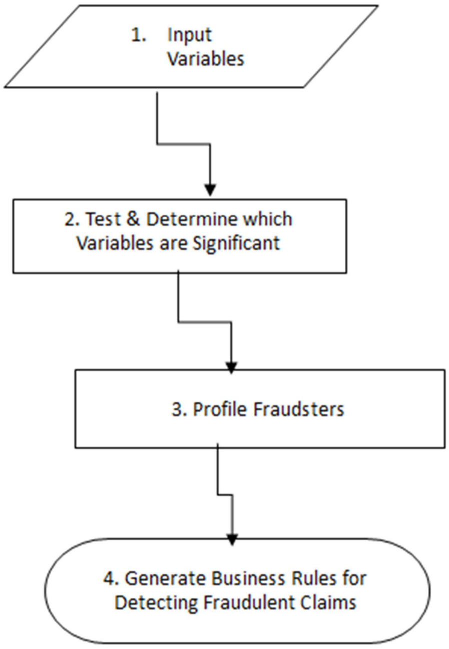
1. Input Variables
2. Test & Determine which Variables are Significant
3. Profile Fraudsters
4. Generate Business Rules for Detecting Fraudulent Claims

Figure 1. Framework for Fraud Detection.

Based on the framework, our first step is to test the significance of variables for identifying fraud. Once we have identified the significant variables, we may use these variables to profile fraudsters and derive the business rules for detecting fraudulent claims.

4.2. Identification & Usage of Significant Variables

Considering that there are many input variables (31) provided to us in the given data set, it is reasonable to come to a conclusion that there may be some variables who do not have a significant impact on detecting fraud. To help determine which variables were actually significant and may help to detect fraud, two different significant tests, the Chi-Square test and the independent samples t-test were applied to the data using SPSS Statistics 22.

Pearson's Chi-square Test for Independence was used to help determine whether there is an association between any two categorical variables being considered. If the significance value is less than .05 (in case of a 95% confidence interval), the null hypothesis of no association between the two variables is rejected. This will indicate that the categorical variable is associated with fraud and vice-versa. Furthermore, one of the assumptions of this test is that there are less than 20% of the cells with a count less than 5. In this case, we look at the Pearson Chi-Square metric to determine the p-value. However, if this assumption is violated, we consider the p-value of the Likelihood ratio metric instead.

The Independent samples t-test was used for comparing the means of the continuous variable for the two independent groups, fraud and non-fraud. If there is a significant difference between the means for fraudulent claims versus non-fraudulent claims, the p-value will be less than .05 - implying that the continuous variable contributes to detecting fraud. The significant variables will be used to assist us in profiling fraudsters versus non-fraudsters and in determining the business rules.

5. Analysis of Results

5.1. Significant Variables

For significance testing of variables, the Chi-Square test used for testing association and the Independent Samples T-Test used for testing differences between means was performed. There are 31 variables (30 categorical and 1 continuous) being considered for identifying fraudulent claims. Table 1, below provides a list of the 31 variables used.

Table 1. Test of Significance of Variables (* denotes statistical significance where p<0.05).

Test of Significance of Variables			
Month*	**Week Of Month Claimed**	**Age***	**Rep Number**
Day Of Week	Make Fin*	Fault*	Deductible*
Week Of Month*	Accident Area*	Policy Type*	Driver Rating
Month Claimed	Sex*	Vehicle Category*	Days: Policy-Accident
Day Of Week Claimed	Marital Status	Vehicle Price*	Days: Policy-Claim
Past Number Of Claims*	Age Of Vehicle*	Age PHF in*	Police Report Filed*
Witness Present	Agent Type*	Number Of Suppliments*	Address Change-Claim*
Number Of Cars	Year*	Base Policy*	

The 0.05 level of significance was used to determine whether a variable was significant for detecting fraud or not, i.e. if the p-value < 0.05, we would concluded that the variable was significant. Variables that were significant were denoted with an '*' in table 1 above. Of the 31 variables, 20 variables were identified as significant. These variables played an important role in profiling the fraudulent claims and generating the business rules for detecting fraudulent claims.

5.2. Fraud Group Profile

To profile the fraudsters, we use the analysis results that we obtained in the identification of Significant Variables in section (5.1) above. Our starting point was to further look into the significant variables and to better understand why there was a significant difference between fraudulent and non-fraudulent claims for the twenty significant variables. Below is a summary of the key findings from the data analysis and significant variables:

Demographic Characteristics of the Fraud Group:

- *Accident Area:* Fraudulent claims tend to mostly occur in urban areas
- *Sex:* Males tend to perform fraud far more commonly than their female counterparts
- *Year:* Fraudulent claims ten to take place within the first two years rather than later.
- *Age of Driver:* Younger drivers(less or equal to 36 years) tend to be more likely fraudulent than older drivers

- *Address Change:* Policy holders who have had their address changed are more likely to be fraudulent
- *Fault:* Policy holder tends to be more likely to be at fault & fraudulent
- *Number of Supplements:* Fraudulent claims are more likely to have no supplements
- *Police Report Filed:* Fraudulent claims tend to have no police report filed.

Vehicle Type of the Fraud Group:

- *Make:* Vehicles made by Pontiac, Toyota and Honda tend to have the most fraudulent claims.
- *Age of Vehicle:* People owning older vehicles (vehicles aged 5 years and more) tend to make fraudulent claims more frequently.
- *Vehicle Category:* Claims involving sedan vehicles are more likely to be fraudulent.
- *Vehicle Price:* Vehicles of low value (priced under 30,000) are more likely to have fraudulent claims filed.

Policy Type of the Fraud Group:

- *Policy Type:* The fraudulent claims is more likely to be
- *Base Policy:* The fraudulent claims are more likely to be a Collision or All Perils type than Liabilities.
- *Agent Type:* The fraudulent claim is more likely to be handled by an external agent.

Claim Characteristics of the Fraud Group:

- *Past number of Claims:* Fraudulent claims tend to have a history of 2 to 4 past claims
- *Deductible:* Fraudulent claims with deductibles amounting to 400 are much more common than those found amongst any other denominations
- *Week of Month Claimed:* Most fraudulent claims tend to be made during the middle of the month
- *Month:* Most accidents tend to occur in January, March, June, July, October or December.
- *Month Claimed:* Fraudulent claims are more likely to be claimed in the months of January, May, October and November

The above twenty characteristics help to provide us with the profile of the fraudulent group. Below, in section 5.3, we derive the twenty business rules to help identify fraudulent claims.

5.3. Derived Business Rules for Fraud Detection

After careful analysis of the data and key variables for fraud identification, we can summarize the results of our analysis and insights found in the twenty business rules derived below. We recommend that these twenty rules be applied to all future claims in the following manner:

5.3.1. Derived Business Rules Used to Test Whether the Claimant Has the ‘Demographic’ Characteristics of a Fraud Profile’

The following four derived rules are highly likely to profile the Fraudster in terms of their demographic characteristics:

- Is the claimant a ‘Male’? If yes, assign the claimant a score of ‘1’, else assign a score of ‘0’.
- Is the driver ‘less than or equal to 36 years of age’? If yes, assign the claimant a score of ‘1’, else assign a score of ‘0’.
- Has the policy holder ever changed their ‘Address’? If yes, assign the claimant a score of ‘1’, else assign a score of ‘0’.
- Was the accident the policy holders ‘fault’? If yes, assign the claimant a score of ‘1’, else assign a score of ‘0’.

At this stage, a claimant may have a minimum score of zero and a maximum score of 4. It is recommended that a claimant with a maximum score of 4, should be immediately processed with the derived business rules for scoring the claimant on the ‘Claim Characteristics’. The derived business rules for the ‘claim’ characteristics of the fraud group is outlined below.

5.3.2. Derived Business Rules Used to Test Whether the Claimant Has the ’Claim Characteristics’ of a Fraud Profile

The following ten rules were derived to help identify whether the claimant is likely to be a Fraudster in terms of their claim characteristics:

- Did the accident occur in either the ‘month’ of January, March, June, July, October or December? If yes, assign the claimant a score of ‘1’, else assign a score of ‘0’.
- Was the ‘month claimed’ either in January, May, October or November? If yes, assign the claimant a score of ‘1’, else assign a score of ‘0’.
- Did the ‘week of month claimed’ occur in the middle of the month? If yes, assign the claimant a score of ‘1’, else assign a score of ‘0’.
- Has the claimant made 2 to 4 ‘past number of claims’? If yes, assign the claimant a score of ‘1’, else assign a score of ‘0’.
- Is the claim within the first two ‘years’ of the policy? If

yes, assign the claimant a score of '1', else assign a score of '0'.

- Was the claim 'deductible' amount $400? If yes, assign the claimant a score of '1', else assign a score of '0'.
- Was the 'accident area' an urban area? If yes, assign the claimant a score of '1', else assign a score of '0'.
- Did the claim have 'no supplements'? If yes, assign the claimant a score of '1', else assign a score of '0'.
- Did the claim have 'no police report filed'? If yes, assign the claimant a score of '1', else assign a score of '0'.
- Was the claim handled by an external 'Agent Type'? If yes, assign the claimant a score of '1', else assign a score of '0'.

We may at this stage compute the overall score of each claimant. The minimum overall score may be 0 and the maximum overall score may be 14. The business may determine at which threshold score they would like immediate processing of the claimant's 'Vehicle Type' characteristics. For example, the business may decide that all claimants who have an overall score of greater than 9, should proceed immediately for 'Vehicle Type' scoring. The derived business rules for 'Vehicle Type' is outlined below.

5.3.3. Derived Business Rules Used to Test Whether the Claimant has the 'Vehicle' Characteristics of a Fraud Profile

The following four rules were derived to help identify whether the claimant is likely to be a Fraudster in terms of their vehicle characteristics:

- Is your 'vehicle type' a Pontiac, Toyota or Honda? If yes, assign the claimant a score of '1', else assign a score of '0'.
- Is your 'age of vehicle' 5 or more years old? If yes, assign the claimant a score of '1', else assign a score of '0'.
- Is your 'Vehicle Category' a sedan? If yes, assign the claimant a score of '1', else assign a score of '0'.
- Is your 'vehicle price' under $30 000? If yes, assign the claimant a score of '1', else assign a score of '0'.

At this stage, the claimant may have a minimum score of 0 and a maximum score of 18. The business may determine at which threshold score they would like immediate processing of the claimant's 'Policy Type' characteristics. For example, the business may decide that all claimants who have an overall score of greater than 13, should proceed immediately for 'Policy Type' scoring. The derived business rules for 'Policy Type' is outlined below.

5.3.4. Derived Business Rules Used to Test Whether the Claimant has the 'Policy Type' Characteristics of a Fraud Profile

The following four rules were derived to help identify whether the claimant is likely to be a Fraudster in terms of their policy type characteristics:

- Is your 'policy type' sedan-all perils or sedan – collision? If yes, assign the claimant a score of '1', else assign a score of '0'.
- Is your 'base policy' all-perils or collision? If yes, assign the claimant a score of '1', else assign a score of '0'.

At this stage, the claimant may have a minimum score of 0 and a maximum score of 20. The business may determine that all claimants who have a final overall score greater than 16, immediate processing of the claim and full details and conclusive evidence should be sought as to whether the claim is fraudulent or not. While claimants with scores greater than ten and less than sixteen should be flagged and further investigated if time and resources allow. A summary of the derived business rules may be seen in figure 2 below. Note that figure 2 is step 4 of figure 1 and that the details of the different computation scores are outlined in sections 5.3.1 – 5.3.5.

Businesses are strongly recommended to deploy their derived business rules in a systematic way so as to ensure consistency in their methodology of identifying and detecting fraud. It is to be noted that over time, the derived business rules can be modified and improved as new information becomes available.

6. Conclusion

Insurance companies are realizing the importance of data analytics in the fraud detection space and are hurriedly opting for expensive fraud solutions that are not aligned to the company's weakness and strengths. In order to leverage data analytics solutions to the fullest, insurance companies should use simple data analytic techniques such as statistical significance testing, then profiling of fraudulent claims by which business rules may be derived, after which, a framework can be built, similar to the one presented in this paper. This paper demonstrated how key variables such as the demographics of the claimant, the claim characteristics, policy and vehicle type may be used to easily identify fraudulent claims and help the business in being more efficient by focusing on fewer variables (in this case 20 instead of 31, a reduction of 35% in the number of variables) and thereby saving time and costs for the business. This is a great help in fraud investigation as more time and focus can be made focusing on the significant variables using the derived business rules. It is hoped that the fraud detection framework presented in this paper will improve the fraud investigation efficiency by reducing the fraud investigation time and costs.

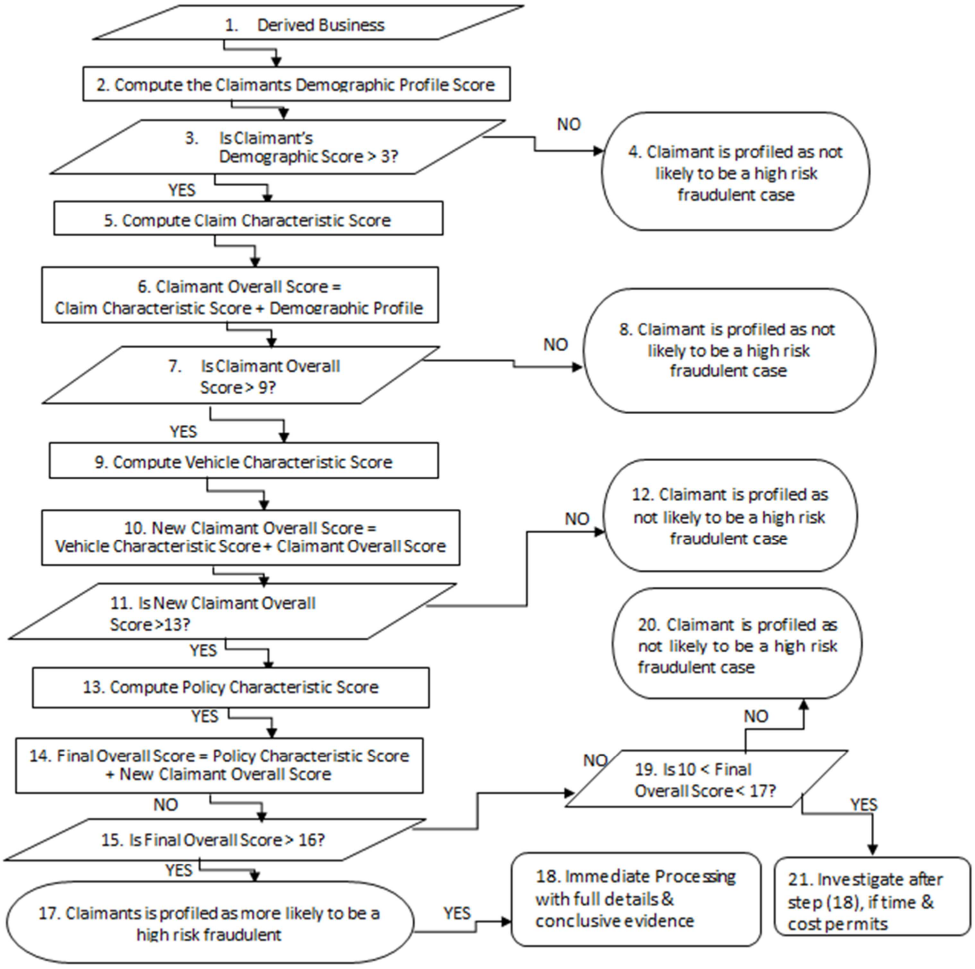
1. Derived Business
2. Compute the Claimants Demographic Profile Score
3. Is Claimant's Demographic Score > 3?
NO
4. Claimant is profiled as not likely to be a high risk fraudulent case
YES
5. Compute Claim Characteristic Score
6. Claimant Overall Score = Claim Characteristic Score + Demographic Profile
7. Is Claimant Overall Score > 9?
NO
8. Claimant is profiled as not likely to be a high risk fraudulent case
YES
9. Compute Vehicle Characteristic Score
10. New Claimant Overall Score = Vehicle Characteristic Score + Claimant Overall Score
11. Is New Claimant Overall Score >13?
NO
12. Claimant is profiled as not likely to be a high risk fraudulent case
YES
13. Compute Policy Characteristic Score
20. Claimant is profiled as not likely to be a high risk fraudulent case
YES
14. Final Overall Score = Policy Characteristic Score + New Claimant Overall Score
NO
NO
19. Is 10 < Final Overall Score < 17?
YES
NO
15. Is Final Overall Score > 16?
YES
17. Claimants is profiled as more likely to be a high risk fraudulent
YES
18. Immediate Processing with full details & conclusive evidence
21. Investigate after step (18), if time & cost permits

Figure 2. Derived Business Rules for Fraud Detection.

References

[1] Bolton, R.J. & Hand, D.J. (2002). “Statistical Fraud Detection: A Review”, Statistical Science, Vol. 17. No. 3, 235-255.

[2] http://www.oxforddictionaries.com/definition/english/fraud.

[3] Wilson, J.H (2009). “An Analytics Approach to Detecting Insurance Fraud using Logistic Regression. Journal of Finance and Accountancy. Vol. 1 Page 1.

[4] Artis, M., & Mercedes, A., & Montserrant, G. (2002). Detection of Automobile Insurance Fraud with Discrete Choice Models and Missclassified Claims. The journal of Risk and Insurance.

[5] Verma, R. & Sathyan, R. M. “Using Analytics for Insurance Fraud Detection: 3 innovative methods and a 10-step approach to kick start your initiative”. Digital Transformation. Pages 1-10.

[6] KPMG International (2006). “Fraud Risk Management: Developing a Strategy for Prevention, Detection and Response”.

[7] Berry, M. and Linoff, G (2000). Mastering Data Mining: The art and science of customer relationship management. John Wiley & Sons, New York, USA.

[8] Costons, M (2010). “Analytics and Claim Fraud: Assembling the proper toolbox to prevent and detect scams”. Claims Magazine. Page 43 – 45.

[9] Spathis, C. (2002). “Detecting falsified financial statements using published data: some evidence from Greece “. Managerial Auditing Journal, 17 (40, 179 – 191.

[10] Spathis, C., Doumpos, M., & Zopounidis, C. (2002). Detecting falsified financial statements: a comparative study using multicriteria analysis and multivariate statistical techniques. The European Accounting Review, 11 (3), 509 - 535.

[11] Phua, C., Lee, V., Smith, K. & Gayler, R. (2005). A comprehensive survey of data mining-based fraud detection research, Artificial Intelligence Review (2005) 1–14.

[12] www.angoss.com/.