A.A.S. **2**, II, 271-288 (2007)

AUTOMATIC FRAUD DETECTION — DOES IT WORK?

By I. Lægreid

ABSTRACT

The aim of automatic fraud detection is to assist claims handlers by selecting claims which are potentially fraudulent. Such methods must be based on information routinely available. The present paper makes use of logistic regression and the validity of these models as demonstrated on fresh data from household and motor insurance. A second contribution is a strategy for selecting claims that should receive further attention.

KEYWORDS

Insurance Fraud; Fraudulent Behaviour; Automatic Fraud Detection; Probability of Fraud

CONTACT ADDRESS

Irene Lægreid, White Label Insurance AS, Harbitzalléen ZA, Box 434 Skøyen, 0213 Oslo, Norway. E-mail: Irene.Lagreid@whitelabel.no

1. Introduction

The paper demonstrates how automatic fraud detection models based on statistical classification techniques are created, and establishes how they succeed in classifying claims as fraudulent. A probabilistic model which links the frequency of fraud to certain claims is used, with explanatory variables, such as age, date of accident and reporting, description of the claim, etc. We discuss the application of the model by means of authentic insurance examples and data from the mass branches of household and motor insurance. However, personal damage and legal aid are excluded. A strategy for when to select a claim for further investigation is discussed and a threshold established. The models are validated empirically by calculating the rate of success as the share of claims which are rightfully captured by the model. In addition, the Hosmer–Lemeshow goodness-of-fit test is used. An automatic fraud detection model is to be viewed as a supplement to the evaluation done by the claims handler in the decision-making process, and is meant to be effected immediately when a claim is reported. However, it is not possible to replace the intuition of the claims handler by a detection model.

There is some literature on fraud detection. Artis *et al.* (1999, 2002) utilise logistic regression in their work in classifying fraudulent claims in

https://doi.org/10.1017/S1748499500000361 Published online by Cambridge University Press

motor insurance. The variables used give characteristics of the insured, the claimant or the policy, in addition to characteristics of the vehicle and the accident. Similar work by Brockett *et al.* (2002) introduced 'supervised learning', based on a correct classification for a respondent group, by using discriminant analysis and Pridit score models for motor insurance. In this work, variables which describe the situation of the insured, like medical health, civil status, etc., are used. Derrig (2002) proposes an interactive claim sorting algorithm development flow, which starts with a systematic benchmark of a random sample of claims and then isolates red flags or other sorting characteristics. Further, he proposes a cluster analysis, like Pridit, or fuzzy clustering of potentially conflicting criteria and the utilisation of models based on neural network, generic algorithm, naïve Bayes or logistic regression.

This work is based on logistic regression, and uses examples and results from Lægreid (2003), where models depending on gender are created for household and motor insurance. Fraud is defined as the situation when at least one of three possible outcomes is recorded: notice of intended prosecution; reduced compensation; or refusal due to exaggerated demands. The first group is unambiguous, while the other two groups might be rather difficult to separate because of the evidential foundation.

The fraud phenomenon includes two major categories: the primary type (25%) is the totally fraudulent occurrences; and the secondary type (75%), with exaggerated claim amounts, is the most difficult to reveal.

Designing an automatic fraud detection model is a continuing process in search of new tendencies, because the behaviour of a dishonest person is likely to change over time. It is necessary, constantly, to evaluate new variables, and, alternatively, to exclude others. In addition, information from the environment, such as crime statistics from the police or from Statistics Norway must be taken into consideration.

2. Method Design through Iteration

Given two distinct multi-variate populations, the fraudulent and non-fraudulent individuals, we assume that we have samples from the two populations available. For this classification problem, a decision rule should be evolved. Based on the information concerning each individual, the decision rule assigns the person to the right group, as is done in Kossack (1963).

The methodology of designing a detection model has several elements, as presented in Figure 1. The core is a probabilistic model which predicts the probability of fraud based on historic customer information. In this analysis a multiple logistic regression model is chosen, as further discussed in Section 4. Other classification techniques with linear classifiers, such as naïve Bayes classifier or Fisher's linear discriminant, are possible to use. Techniques like

https://doi.org/10.1017/S1748499500000361 Published online by Cambridge University Press

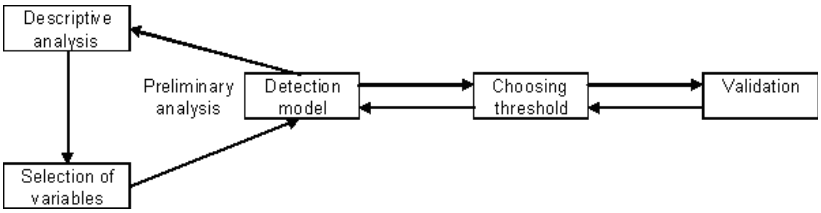


Figure 1. The process of creating detection models

k-nearest neighbour, boosting, decision trees, neural or Bayesian networks or hidden Markov models are also possible.

The data are the main challenge. A detection model which is meant to be used immediately on reported claims requires explanatory variables based on easily accessible data. This requires the claims handler to use information which is available immediately. More and more claims are settled while the customer is on the phone, so the claims handler has to make a quick decision whether to pay right away or to ask for more information. The time aspect is of great importance. Designing detection models is, indeed, a compromise between how many variables to request and the pressure to reach a quick answer. Moreover, due to restrictions from legislation and the protection of privacy, the desired data are not always available. How to organise the data as a training data set is further discussed in Section 3.

In the preliminary analysis (Section 5) one has to search for explanatory variables which, in fact, describe the person and his behaviour when a claim is reported. Here the simple strategy of fitting single covariates is used. At this stage, an iterative procedure, as indicated in Figure 1, is beneficial. It is important to be able to step forwards and backwards in the process, because the information concealed in the data will appear as the analysis progresses. The most labour-intensive part is indeed the preliminary analysis, because it is not obvious which variable does explain fraud and which does not. Further, examples from various findings in preliminary analyses for household and motor insurance are presented.

Selecting a decision rule to be used in the classification of an individual is indeed a major key to the efficient use of automatic detection models, i.e. selecting the 'right' claim for further investigation. In Section 6, how to choose a threshold for the logistic regression models is discussed and exemplified by empirical research on updated test data sets from motor and household insurance. In addition, the differences between the two branches are discussed.

Validation is the final step of the process. Section 6 presents empirical results from the validation of a household model on test data from June 2005. To measure how successful the detection model is in capturing the

https://doi.org/10.1017/S1748499500000361 Published online by Cambridge University Press

fraudulent claims, a frequency (*rate of success*) is worked out on percentile intervals based on sorted probabilities of fraud. In addition, a more formal way of testing the model is derived using the Hosmer–Lemeshow goodness-of-fit test (Hosmer & Lemeshow, 1989).

It is possible to use the framework presented in Figure 1 for different lines of business, such as travel or liability insurance. It is also possible to adapt to certain areas of life insurance such as group life. Designing detection models to be used in commercial branches requires a certain adjustment of the variables, because the insured is usually a company or an organisation.

3. Organising the Data

Models depending on gender and line of business are used in the analysis. The proposed list below suggests explanatory variables which might be implemented in the models. When a claim is reported, the usual variables recorded could be summarised in six relevant groups. The list of variables is not entirely complete for all possible types of accidents:

(1) *biographical data:* the age of the claimant when the claim was reported; sex;
(2) *geographical data*: the place of residence: like county; municipality; urban or not urban; description of the geographical location where the incident took place;
(3) *data describing the line of business:* which product or product group;
(4) *time-based data — the occurrence:* when the accident/injury happened — year/month/weekday/moment of time;
(5) *time-based data — the moment of reporting:* when the costumer reported the accident/injury; the reporting date (the first date the company was given notice of the occurrence); and the date of taking out the policy; and
(6) *data based on codes of analysis:* a description of how the incident happened, like: specification of the methods (theft, fire); the source of fire (by lightning, intentionally, unknown); a description of the security level (the presence of an alarm).

In addition, the fraudulent claims should be supplied with the claim status in accordance with the definition of fraud given in Section 1. Finally, all the adequate variables should be put together in one record per claim as the foundation of the analysis. These variables are regarded as neutral characteristics. Variables like payment remarks, nationality, or whether the customer has a clean record, are considered as sensitive under Norwegian legislation, and we are not allowed to use this information. There is a register (FOSS) of all claims reported in Norway, but the insurer is not permitted to use this information for analysis, and the insurer is limited to using only its

https://doi.org/10.1017/S1748499500000361 Published online by Cambridge University Press

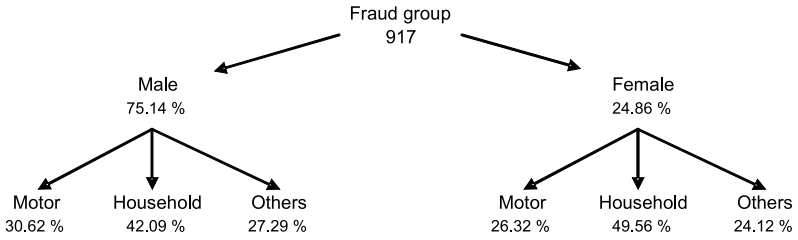


Figure 2. The fraud group

own data. In motor insurance, it is very common for customers to change insurers frequently. Thus, we have chosen not to use the number of claims reported to the insurer as an explanatory variable. Thus, the number of claims reported to one company is only a part of the reality, and is chosen not to be used as an explanatory variable in the paper.

One has to ensure that the data used from previous periods provide a correct picture of the present situation and if the explanatory variables are relevant in a detection model which is meant to be used on future claims.

To create a training data set, the historic data were divided into two groups: the fraud group; and the control group. The fraud group consisted of all the identified occurrences of fraud in a certain predefined period, and the control group represented all the other claims recorded in the same period. The fraud group was based on 917 claims, of which 689 were reported by men and 228 by women. Figure 2 illustrates how the data in the fraud group were organised before executing the analysis which is illustrated here.

The data were divided by sex, because the analyses indicated, at an early stage, that behaviour was significantly different between women and men. This was a persistent feature of the data. In the fraud group were cases where the claimant was reported to the police, or where the claim was refused or reduced by the insurer.

The control group consisted of all the other reported claims from the same period as the identified fraudulent claims. Another approach is to have a fraud group and a control group which are equal in size. This approach has been used by Artis *et al.* (2002). Based on a predefined fraud group of 1,000 observations and a control group of the same size, indicators of fraud are identified by a logistic regression model.

Here the control group consisted of 644,939 claims, 423,976 male and 220,963 female. Figure 3 illustrates how the data were organised before carrying out the analysis.

https://doi.org/10.1017/S1748499500000361 Published online by Cambridge University Press

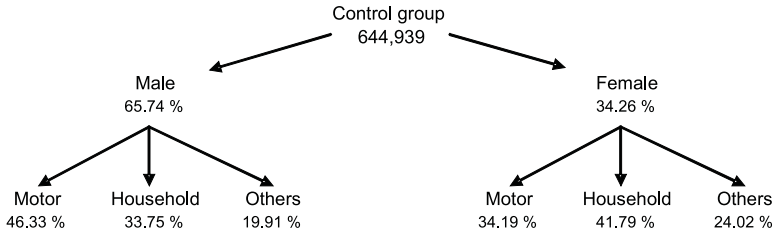


Figure 3. The control group

4. Probabilistic and Empirical Modelling

Claims are recorded as fraudulent or not, with $p_i = p(x_i)$ being the probability of fraud, given a vector x_i of explanatory variables associated with claim i, $p_i \in [0, 1]$.

For k explanatory variables and $i = 1, \ldots, n$ individual claims, the multiple logistic regression model is defined as:

$$\log(p_i/(1 - p_i)) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} = \beta' x_i$$

with β being the unknown parameter vector to be estimated.

To estimate β, one can use the method of maximum likelihood. Assuming the responses y_i (the claims in the database) to be independent and $\Pr(y_i) = p_i^{y_i}(1 - p_i)^{1-y_i}$, the maximum likelihood estimate of β is found from maximising the loglikelihood $\log L(\beta)$ with respect to β. This is worked out by using the Genmod Procedure in SAS (1999).

5. Data Analysis

5.1 *Examples from a Preliminary Analysis for Household and Motor Insurance*

The behaviour when reporting a claim to the company can expose the intention of the customer. How the customer expresses himself and what information he gives and does not give provides valuable knowledge. Once a phenomenon or a tendency is revealed or indicated in the material, the variable is marked as interesting, and should be examined further.

One way of starting is to look at the distribution of the number of claims in the fraud group versus the control group. The two groups from the example in Figure 4 show different curves. This difference might be captured by an explanatory variable. Examples of the phenomenon presented in Figures 4 and 5 are from Lægreid (2003). The data were recorded from 1993 to the end of 2001 by the claims handlers, and classified by the investigation department.

https://doi.org/10.1017/S1748499500000361 Published online by Cambridge University Press

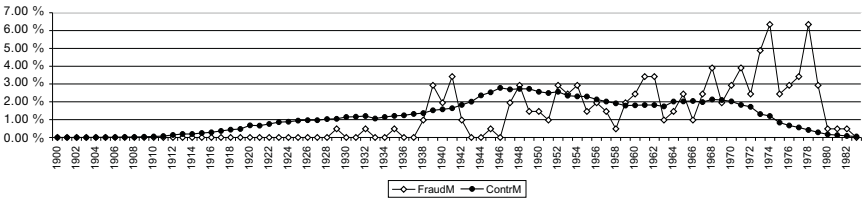


Figure 4. The number of reported motor claims (male) by year of birth in the fraud group versus the control group, as at 31 December 2001

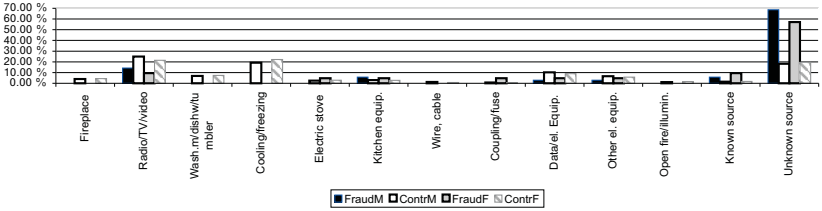


Figure 5. The number of reported claims for the household line, for the fraud group and for the control group, by different sources of fire, as at 31 December 2001

An interesting phenomenon which is only present in the fraud group is the dramatic increase in the level of occurrences, i.e. fraudulent claims among those with years of birth after 1968.

Figure 5 illustrates another example of a phenomenon where the sources of fire are reported as 'unknown'. The columns show that the distribution within groups indicates a higher level of reported unknown sources among the dishonest than among the others. This is a phenomenon which is well known by investigators. Eager not to be revealed, the customer tends to say as little as possible.

Seasonal variations in the accident date and the reporting date are easy to check. Figure 6 shows the household (female) graph for the number of claims reported from 1993 to 2001 in different months. The distribution shows how the reporting pattern in the fraud group differs from the control group. In January and February the occurrences of fraud amongst females seem suspiciously high, and the same tendency does not appear in the control group. This phenomenon could indicate that women are short of money after Christmas, and that checks should, perhaps, be especially focused on this period. The phenomenon has to be verified as an explanatory variable in the models later on.

https://doi.org/10.1017/S1748499500000361 Published online by Cambridge University Press

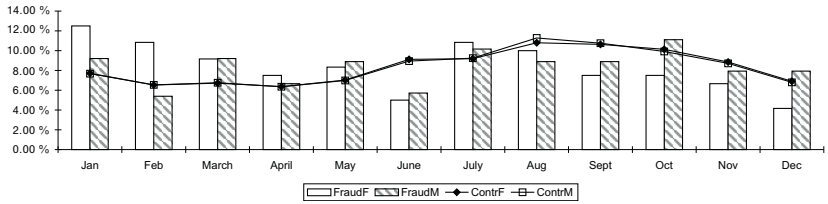


Figure 6. The number of reported claims distributed by reporting month — household (female)

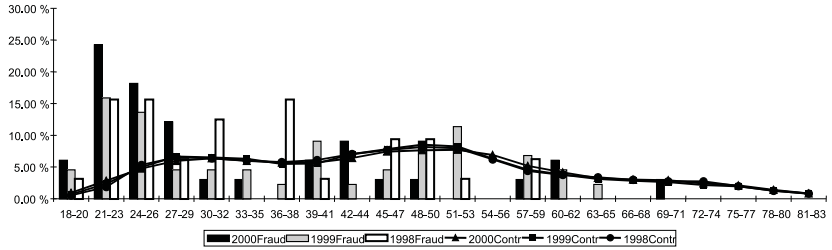


Figure 7. The number of reported claims distributed by age groups — private motor (male)

In the preliminary phase one might also look for indications on how to group the variables efficiently. Numeric variables with finite outcomes, such as age, might end up as 'not significant' if not grouped suitably. Variables described in the text as a type of accident could be grouped efficiently by using a binary outcome as 'theft' or 'not'.

Figure 7, for private motor (male), shows the situation between the distributions of number of the claims within the fraud group and the control group. The claims are divided into different age groups for those reported in 2000, 1999 and 1998.

The curves show graphically how to group the variable 'age' efficiently. The pattern indicates that younger age groups, perhaps, should be grouped more finely than age groups above 50, etc., due to the number of fraud cases. In fact, the curve reveals that younger customers show a tendency of a much higher frequency of fraud, and, thereby, are interesting to examine more closely.

The aspect of time is essential, as illustrated by the following example. How to group the variable 'delay', which is defined as the interval between the alleged accident date and the time of reporting, is shown in Figure 8.

https://doi.org/10.1017/S1748499500000361 Published online by Cambridge University Press

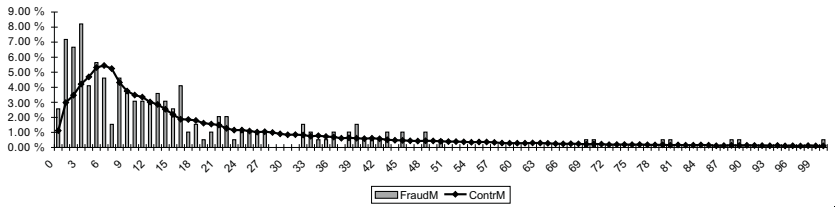


Figure 8. The number of reported claims distributed by days of delay from accident to reported date — private motor (male)

Normally the claims are registered into the claims system right away, or at least within a couple of days. This provides information on how the dishonest customer plans to act when he or she knows that the claim is fraudulent.

The distribution within groups for the variable 'delay' is plotted. By comparing the two distributions, it seems as if the curves are different. There appears to be a shorter period of time from the incurred date to the reporting date in the fraud group. Logically, if the occurrence is claimed to have taken place some time ago, it would be more complicated to remember all the details. To seem convincing could obviously be a challenge. Moreover, the fraud curve has peaks which are interesting to examine. They indicate that the variable should be classified with smaller intervals in the first three months after the declared accident day than is necessary later on. The tail of the fraud curve terminates after 100 days, in comparison with the curve of the control group which continues.

5.2 *Selecting Explanatory Variables*

There are many potentially useful covariates for explanatory variables, and a preliminary search for the good ones is necessary. Different approaches may be developed. Here the simple strategy of fitting single covariates has been used.

The following example presents a model with one covariate which is the age group. The plot in Figure 9 shows the estimated parameters for each age group. The lowest β_1 value (min β_1) could be interpreted as the most interesting group, and the other groups ranked from this starting point. The right axis represents the p-value of the χ^2 test of significance of the regression parameter. A reference line is drawn at the 5% level. The significant age groups are below the line. The interpretation of the plot is that the relative size of the coefficients gives a picture of which age group to focus on as the most important. The age group 21 to 23 has the lowest estimated value, the next is the age group 24 to 26, etc. The phenomenon of young men with fraudulent behaviour was apparent from Figure 7.

https://doi.org/10.1017/S1748499500000361 Published online by Cambridge University Press

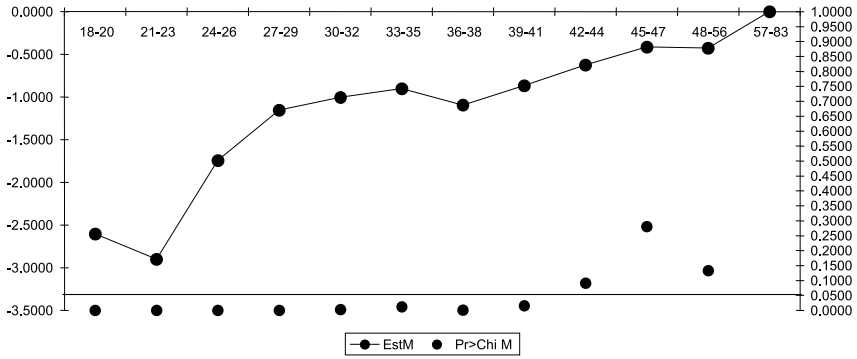


Figure 9. The estimated parameters for each age group for motor insurance from a logistic regression model with one covariate

Another example is the regression model with one covariate corresponding to the situation in Figure 8. The estimated coefficients in β_1 are shown in Figure 10.

By interpreting the figure, it is obvious that the customer tends to report the fraudulent claim within five days of the alleged accident date. Moreover, the figure indicates that the male most likely reports the occurrence after approximately two weeks, three weeks, one month or 1.5 months. Interestingly, the chosen date to get in contact with the company might not be random. The figure indicates that a person who deliberately reports a claim which is invented or has an exaggerated demand tries to plan the 'right day' carefully

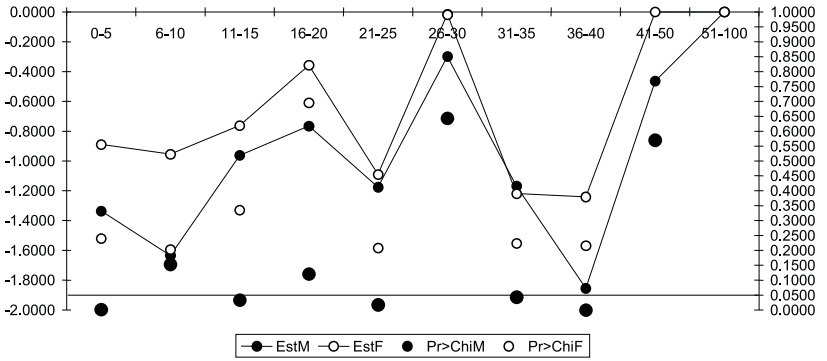


Figure 10. The estimated parameters for each interval of delay for motor insurance from a logistic regression model with one covariate

https://doi.org/10.1017/S1748499500000361 Published online by Cambridge University Press

to make contact with the company. Anxious not to be revealed, he or she wants to appear as anyone else. However, the distribution of the number of claims in the fraud group reveals the dishonest behaviour.

6. Thresholding and Validation

The formal method is to pronounce a case i fraudulent if the probability of fraud $p_i \geq P^*$, where P^* is a threshold. Hence, all claims i with the following condition are to be chosen: $\{i \mid p_i \geq P^*\}$.

How we choose P^* is a compromise between false 'negatives' and false 'positives', i.e. we would like to include as many frauds as possible among cases selected and yet avoid the cost of innocent people being scrutinised. An 'optimal' selection of P^* is only achieved by including expenses on each decision which lead us to decision theory, as in Artis *et al.* (1999).

Another possibility is to adapt the threshold to recent experience, for example that of the previous year. In this paper, the reporting month is included amongst the explanatory variables and the solution presented looks back exactly 12 months.

Suppose that we are at the start of month k. One year earlier we have registered probabilities of fraud $p'_1, \ldots, p'_n$, where n is the member of observations in month k. The marked p_is refer to the probabilities from the test data set. When sorted in descending order $p'_{(1)} \geq \ldots \geq p'_{(n)}$, our proposal for the current month k is to pronounce case i as suspicious if its probability of fraud p'_i satisfies $p'_i \geq P^*$, where $P^* = P_{(n\varepsilon)}$, and ε is the chosen percentage level of claims stopped by the detection model.

P^* varies over the year, and is automatically revised every year. In detail there are many ways to implement this idea. It seems worthwhile to let ε vary with the line of business.

The detection models described in the paper have been formally validated on fresh data from motor (male) and household (male). The test data include all claims registered in June 2005.

The sorted list $p'_{(1)} \geq \ldots \geq p'_{(n)}$ was presented to the investigators for further inspection, starting with the motor test data and the highest probability of fraud. The claims were marked as either 'interesting' or 'not', without going into further details. The 'interesting' claims should have been looked into twice before settlement in accordance with the chosen definition of fraud. Then, it was possible to determine a suggested threshold P^* empirically. The results indicated that the 5% percentile appeared to be a credible guideline for motor. Ideally, one should examine every one of the claims thoroughly, which is rather time consuming, as one has to check out (often) hand-written claims forms, police reports, whether the customer has payment remarks or undesirable behaviour towards the company, etc. Nevertheless, this is done for household, as shown in Figure 12.

https://doi.org/10.1017/S1748499500000361 Published online by Cambridge University Press

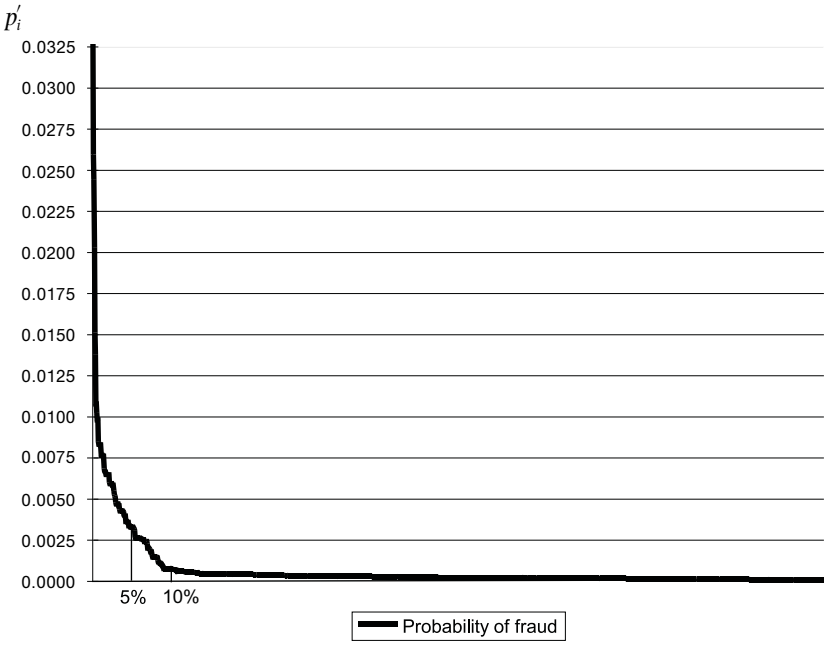


Figure 11. The sorted probabilities of fraud for motor (male)

Figure 11 presents the p_i's of motor (male) for all n claims reported in the month of June 2005. If the probability of fraud p_i' satisfies $p_i' \geq P^*$ and $P^* = 0.0034$, this corresponds to the 'worst' 5% of the claims, hence $\varepsilon = 5\%$. If $\varepsilon = 10\%$, then $P^* = 0.0008$.

Before presenting the sorted list of p_i's for household to the investigators, a graph corresponding to Figure 11 was made, see Figure 12.

For household, $P^* = 0.0032$ when $\varepsilon = 5\%$ and $P^* = 0.0016$ when $\varepsilon = 10\%$.

By comparing the curves in Figures 11 and 12, the rate of decrease appears to be different between the two lines of business. To illustrate the difference, the two plots are combined in a segment of the percentiles in Figure 13, with p_i's shown on the vertical axis and the percentage of claims captured by the detection model on the horizontal axis.

The motor curve has a higher rate of decrease towards zero than the household curve. This might indicate greater uncertainty in household, especially in the segment between the 5% percentile and the 25% percentile. This phenomenon could indicate a gradual development in household from 'the most fraudulent claims' below 5% and 'the possible fraudulent claims' between 6% and 20% to 25%. The threshold for household is anticipated to

https://doi.org/10.1017/S1748499500000361 Published online by Cambridge University Press

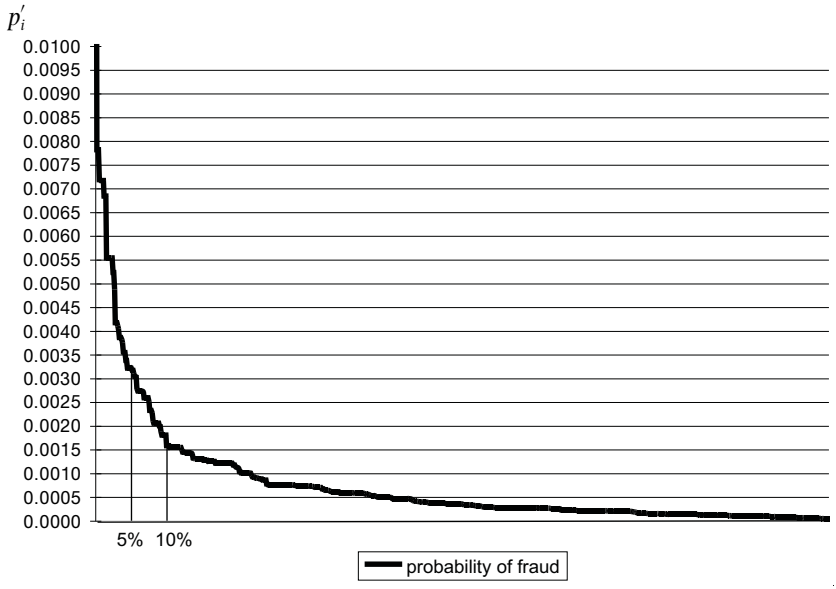


Figure 12. The sorted probabilities of fraud for household (male)

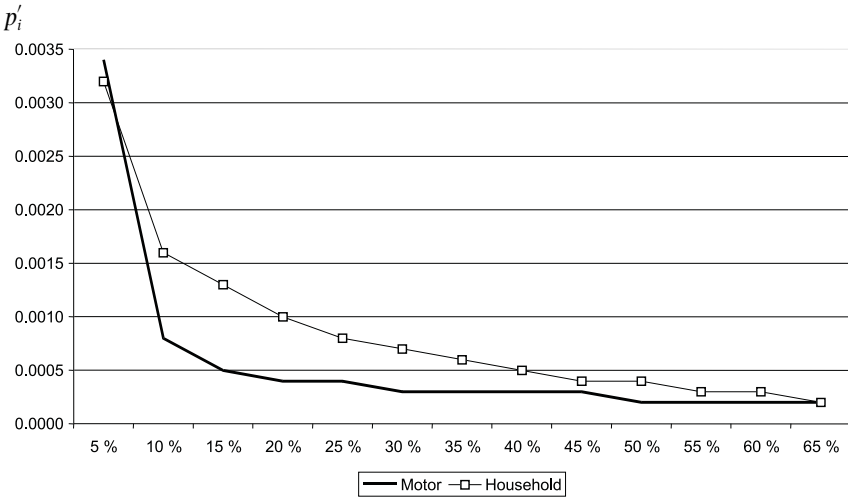


Figure 13. The sorted probabilities of fraud for household (male) and motor (male) plotted against the percentage level of claims stopped by the detection models

https://doi.org/10.1017/S1748499500000361 Published online by Cambridge University Press

be higher than for motor, because the evidence is easier to produce. In addition, the proportion of secondary fraud is significantly higher in household.

To validate the household model, the sorted list $p'_{(1)} \geq \ldots \geq p'_{(n)}$, $n = 920$ claims, from June 2005 was presented, to the investigators for a thorough and extensive inspection. The sorted p'_is were placed into intervals of percentiles. Let ξ denote the different segments of percentiles, such as:

$$\xi \in \{[0, 0.05), [0.05, 0.10), [0.10, 0.15), [0.15, 0.20), [0.20, 0.25), \ldots, [0.95, 1.00)\}.$$

Let N_ξ be the total number of claims in interval ξ, $i \in [1, \ldots, N_\xi]$.

Let $\kappa_{i\xi} = 1$ if claim i in segment ξ is fraudulent or zero otherwise.

To measure how successfully the model captures the fraudulent claims in each segment, *the rate of success* within each ξ is defined as:

$$\lambda_\xi = \frac{\sum_{i=1}^{N\xi} \kappa_{i\xi}}{N_\xi}.$$

The segment $\xi = [0, 0.05)$ includes the 5% 'worst', $\xi = [0.05, 0.10)$ the 5% to 10% 'worst', etc. $N_\xi = 46$ in the test data set. All claims captured by the detection model below the 25% percentile were strictly evaluated in collaboration with the investigation department. The definition of fraud covers notice of intended prosecution, reduced compensation and refusal due to exaggerated demands. Thereby, obtaining success or declaring a certain claim 'possibly fraudulent' must be interpreted as 'there is a certain need to further check the claim before settlement'.

The number of claims within each interval which turned out to be fraudulent or dubious:

$$\kappa_\xi = \sum_{i=1}^{N_\xi} \kappa_{i\xi}$$

were $\kappa_{[0.00,0.05)} = 39$, $\kappa_{[0.05,0.10)} = 31$, $\kappa_{[0.10,0.15)} = 15$, $\kappa_{[0.15,0.20)} = 13$, $\kappa_{[0.20,0.25)} = 11$.

Figure 14 presents the different values calculated for λ_ξ for the detection model on the vertical axis and the different segments of the percentiles on the horizontal axis.

The rate of success λ_ξ in the segment $\xi = [0, 0.05)$ is 0.85. This result is to be interpreted as meaning that 85% of the claims need to be looked into twice before settlement because they might be dubious or fraudulent. λ_ξ in the next segment is also surprisingly high, close to the value of 0.67. This empirical result indicates that the threshold P^* for household is presumably twice the size of the limit for motor. The curve drops to half the size between the [0.05, 0.10) and the [0.10, 0.15) segment. Afterwards it shows a slightly

https://doi.org/10.1017/S1748499500000361 Published online by Cambridge University Press

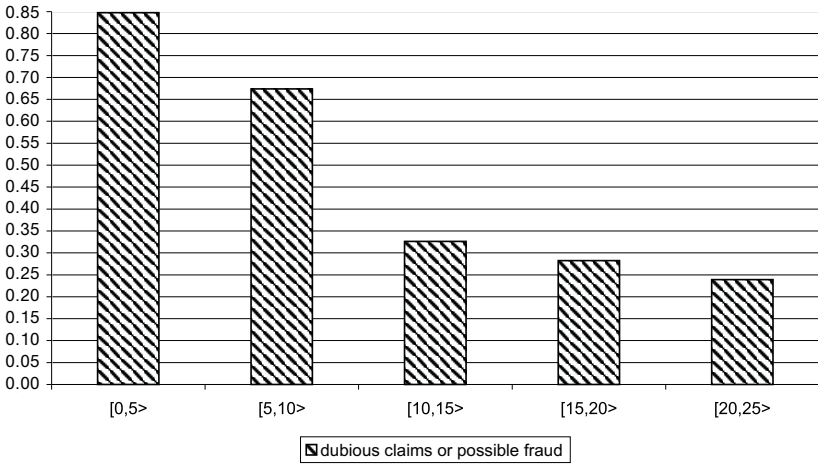


Figure 14. The rate of success λ_ξ for the detection model within each segment of percentiles for household (male)

decreasing tendency. This result indicates that P^* should be at least the 10% percentile for household, and most likely even higher.

A more formal way of testing the model is to use the Hosmer–Lemeshow goodness-of-fit test (Hosmer & Lemeshow, 1989) in the SAS Procedure Procedure Logistic (1999). Hosmer & Lemeshow recommend partitioning the observations into ten equal size groups according to their predicted probabilities, where n_j is the number of observations in the jth group, $O_j = \sum_i y_{ij}$ the observed number of cases in the jth group and $E_j = \sum_i \hat{p}_{ij}$ the expected number of cases in the jth group. The test statistic then becomes:

$$\hat{C} = \sum_{j=1}^{10} \frac{(O_j - E_j)^2}{E_j(1 - E_j/n_j)} \sim \chi_8^2$$

which is χ^2 distributed with eight degrees of freedom.

Before executing the test in 2007, the fraud database from 1993 to 2001 had to be reconstructed due to changes in the data system. It was then possible to identify 150 more cases of fraud, because the national identity numbers were now consistently filled out in the data warehouse.

The household male model from 2001, now executed on reconstructed data, is given in Table 1.

https://doi.org/10.1017/S1748499500000361 Published online by Cambridge University Press

Table 1. Household male model from 2001, using reconstructed data

	No-fraud		Fraud	
	Observed	Expected	Observed	Expected
1	13,421	13,425	176	172
2	13,416	13,418	86	84
3	13,584	13,575	47	56
4	13,531	13,536	46	41
5	13,560	13,557	28	31
6	13,335	13,330	17	22
7	13,012	13,011	14	15
8	13,730	13,728	9	11
9	13,150	13,158	16	8
10	12,579	12,580	6	5

Since the p-value is 9.9%, this model shows no evidence of lack of fit based on the Hosmer–Lemeshow statistic.

To determine the effectiveness of the classification techniques, one might also calculate the probability of misclassification, with $p(i \mid j) =$ the probability of the misclassification of an individual who belongs to population π_i into π_j. Figure 15 presents four different situations which could appear when dealing with fraud detection. The horizontal axis represents the 'truth', in other words what is actually the case, and the vertical axis represents the action.

In situation I it is actually a fraudulent case and one declares it correctly as fraudulent. In situation II one declares a fraud when it is not the case, which corresponds to a Type 1 error. In situation III one does not declare fraud even if it actually is present, which corresponds to a Type 2 error. In situation IV one correctly declares no fraud. The aim is to maximise the probabilities of revealing fraud in situations I and IV and to minimise the probabilities in situations II and III.

If not every single claim is checked in advance, there will always be an intersection between the fraud group and the control group, as indicated in

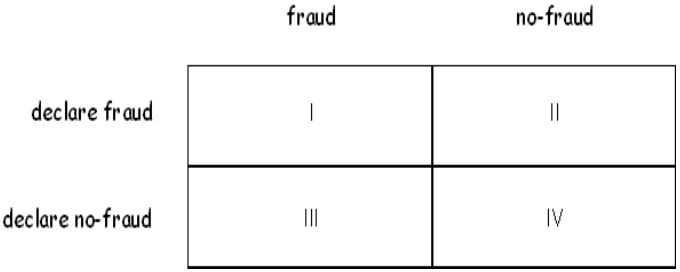


Figure 15. The possible situations

https://doi.org/10.1017/S1748499500000361 Published online by Cambridge University Press

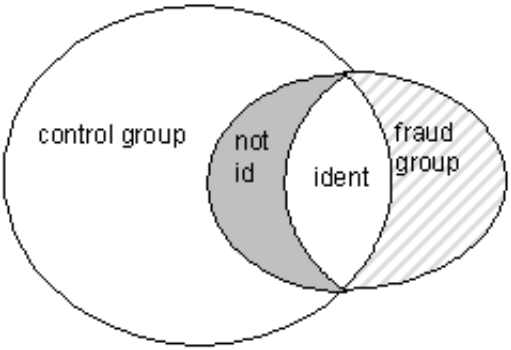


Figure 16. The intersection between the fraud and the control group

Figure 16. The fraud group should consist of the segment marked identified (ident) i.e. the cases which incorrectly appear in the control group because they were not revealed and also the segment not identified (not id).

The 'not id' segment is characterised by unknown criteria and represents a challenge for future analysts. The process of revealing fraud will always be a race between the company and the fraudulent customers. The creativity in inventing new methods of committing fraud seems to be considerable.

8. Concluding Remarks

This paper demonstrates automatic fraud detection models in non-life insurance by using statistical classification techniques as logistic regression, and presents results from validation and implementation on fresh data in household and motor insurance. The main idea is to assist the claims handlers by evaluating the risk of fraud probabilistically by utilising historic customer information and explanatory variables, which should be those available.

The challenge is to ensure the availability of relevant data to assist in characterising fraudulent behaviour among customers and to have an intuition about the future. The behaviour of the clients is changing rather rapidly as they evolve new methods of effecting fraudulent acts, thus the creativity of the analyst is vital. Automatic detection models have to be developed in the specific company on the actual portfolio, and based on the acquired accumulated experience from the same establishment.

How to set a threshold for selecting the right claims is discussed. The functionality of the models is verified by validation results from collaboration with investigators and a frequency (rate of success) on how successfully the

https://doi.org/10.1017/S1748499500000361 Published online by Cambridge University Press

model captures fraudulent claims is worked out. In addition, validation by using Hosmer–Lemeshow goodness-of-fit test is presented.

Acknowledgement

The modelling framework in this paper is based on the thesis 'RBNI — detecting fraud in non-life insurance'. The author is grateful for the comments received from Erik Bølviken, University of Oslo and John Dagsvik at Statistics Norway.

References

Allison, P.D. (1999). Logistic regression — using the SAS-system. SAS Institute Inc., 36-39.

Artis, M., Ayuso, M. & Guillen, M. (1999). Modeling different types of automobile insurance fraud behavior in the Spanish market. *Insurance: Mathematics and Economics*, **24**, 67-81.

Artis, M., Ayuso, M. & Guillen, M. (2002). Detection of automobile insurance fraud with discrete choice models and misclassified claims. *Journal of Risk and Insurance*, **69**(3), 325-340.

Brockett, P.L., Derrig, R.A., Golden, L.L., Levin, A. & Alpert, M. (2002). Fraud classification using principal component analysis of RIDITs. *Journal of Risk and Insurance*, **69**(3), 341-371.

Derrig, R.A. (2002). Insurance fraud. *The Journal of Risk and Insurance*, **69**(3), 271-287.

Hosmer, D.W. & Lemeshow, S. (1989). *Applied logistic regression*. New York: John Wiley & Sons, Inc., 147-156.

Lægreid, I. (2003). RBNI — detecting fraud in non-life insurance. *Cand. Scient. Thesis*, University of Oslo, 1-43.

Kossack, C.F. (1963). Statistical classification techniques. *IBM Systems Journal*, **2**(2), 136-151.

SAS Institute Inc. (1999). The genmod procedure. SAS/STAT user guide, 1368-1437.

https://doi.org/10.1017/S1748499500000361 Published online by Cambridge University Press