"Science Stays True Here"
Journal of Mathematics and Statistical Science, 311-321 | Science Signpost Publishing

A Data Mining Approach: Application to the Extraction of the Characteristics of IARD Products in the Insurance Sector

Nora MOSBAH LOUNICI

High National School of Statistics and Applied Economics ENSSEA, Universities of Kolea, Algeria.

Laboratory of applied statistics

Khadidja SADI

High National School of Statistics and Applied Economics ENSSEA, Universities of Kolea, Algeria.

Laboratory of applied statistics

Nadjib LOUNICI

School of Commercial High Studies

Abstract

In this study, we were interested to multirisk home products (MH) and professional multirisk (MP) of the branch IARD[1] insurance, which covers fire, accidents and various risks of an Algerian insurance company (SAA). We want to know the variables that best characterize each of the two products. To solve this problem, we have combined three methods borrowed from Data Mining and decisional statistics. The use of data mining tools to realize a classification and to bring out the most informative variables is usual tasks of this discipline. The aim of this work is to extract knowledge from an actuarial database by adopting an approach to data mining and to compare the results obtained by querying three techniques, implemented in data mining software: TANAGRA.

Keywords: Data mining, Insurance IARD, Regression, Step Disk, V-test

1. Introduction

The Data Mining [1] is the set of techniques and methods of data analysis, coming from different fields such as database, statistics, artificial intelligence, information theory ...etc.

[1] For property and casualty insurance know in Algeria as IARD

The objective of this discipline is the discovery of trends and correlations hidden from a mass of data. This domain developed in companies is dedicated to the search for strategic information and new knowledge that could surprise the expert and help the decision-maker to improve the performance of his institution.

Otherwise, financial institutions hold gigantic databases, recording socio-demographic or transactional information. These stored data, if they are properly exploited, they can provide real added value for the company. In this study, we were interested in the field of insurance which occupies a crucial role in the economic and social activity, thanks to the benefits offered by the insurance companies.

In the case of the occurrence of risks, the insurance undertakes to compensate the insured by proposing a multitude of products, among others, IARD insurance products (fire, accidents and various risks) [2].

This branch has known a remarkable evolution in the Algerian market. It has recently surpassed car insurance with a weight of more than 40%, dominated mainly by companies and public institutions. This may be caused in part to the consciousness of individuals of the importance of ensuring and protecting themselves against various risks.

Among the products of the IARD branch, we are targeting more particularly the home multirisk (MH) and the professional multirisk (MP) of the Algerian Insurance Company (SAA). The concept is to offer a combination of guarantees such as fire, theft or civil responsibility. The concerned products are destined mainly for professionals and individuals wishing to protect their patrimony by subscribing an insurance contract [3]. This article presents an experimental study consisting, on one hand, on extracting the characteristics of the products MH and MP, and on the other hand, looking for the specific features of each of the two products. To do this, we suggest a generic approach combining three methods borrowed from Data Mining and statistics.

- *the Binary logistic regression* is a statistical model to study the relationship between the explanatory variables and the branch variables we are trying to explain.
- *The Tanagra StepDisk procedure* is used to extract the most pertinent variables.
- And finally, *the criterion of the test-value,* which is useful to characterize each of the groups in order to extract, in a coherent way, the most important attributes.

Those three methods are implemented in a data mining software, which we used to lead our experiments, it is Tanagra software. So, we define the set of data used to lead this study in section 2.

In the third section we apply the binary logistic regression to evaluate the contribution of the explanatory variables to the explanation of the target variable.

Then, we select the most pertinent variables by using the STEPDISC method according to two approaches, ascending and descending. Finally, Section 3.3 constitutes on the extraction of the characteristics of each branch by the V- test tool.

2. Data and Methodology

2.1. Preparation of Data

The data preparation phase is particularly important in the data mining process [4]. It is directly related to the quality of the results. This explains why data cleaning usually requires about 80% of time in a data mining study [5]. We realized our experiments on a dataset of 14 variables and 544 individuals, collected over a period of 4 years by 31 agencies of the Algiers Regional Direction specific to the national insurance company (SAA). (01/01/2010 to 31/12/2013).

Type	Independent variables	Modalities
Qualitative	Legal status	Compagny = 0 Particular = 1
	State	Bonus = 1 Malus = 2 No = 3
	Type of the sinister [2]	DDE = 1 RC = 2 VOL =3 BDG = 4 INC=5
Quantitativ	Disaster year	2010 =1 2011= 2 2012 =3 2013 =4
	Year insurance contract	2010 =1 2011 =2 2012 =3 2013=4
Quantitatif Discrétisé	Premium Insurance	prim1 =]1039.4;7763] ; prim2=]7763; 27749] prime3 = [=]27749 ; 61645] ; prime4 = > 61645
	Sinister Evaluation	Eval1 =]0 ; 40000]; Eval2 =]40000 ; 148770] Eval3 = > 148770
	Sinister Amount	Amount1 =]315 ; 16200] ; Amount2 =]16200 ; 40100] Amount3 = > 40100
Dependent Variable		**Branch** MH=0 ; MP=1

Table 1: the retained variables after pretreatment

[2] DDE (water damage) BDG (breaking of the glass), INC (Fire) et RC (public liability)

After carrying out a univariate and a bi-varied analysis, we decided to keep the most informative explanatory va-riables having a discriminating power compared to the variable to be explained. These variables are presented in (Table 1).

The (Figure 1) shows the distribution of our samples by class. The target variable "Branch" is dichotomous.

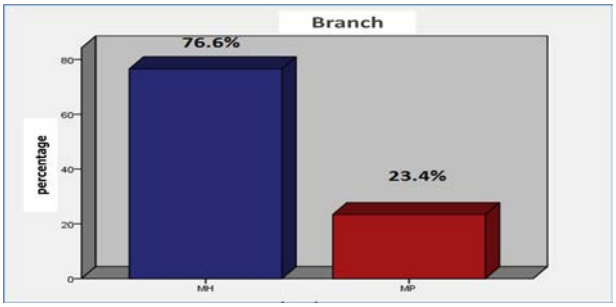


Figure 1. Distribution of Insureds by Branch

We observe that nearly 77% of subscribers are assured of the contract MH. On the other hand, only 23.4% subscribed to an MP. The distribution by Branch and Legal Status is illustrated in Figure 2. It shows that 90% of companies are insured in MP and 83% of individuals in MH, which makes sense since the product MP is intended especially for companies whereas MH concerns the individuals.

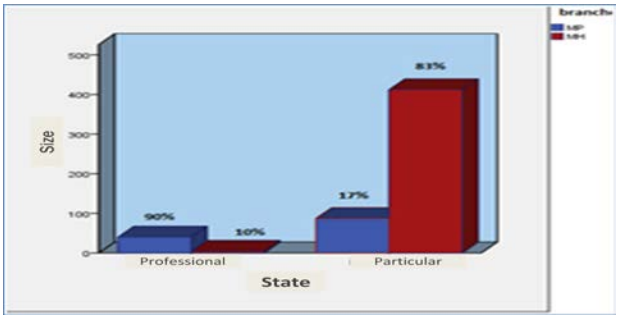


Figure 2. Distribution of Insured by Branch and by Legal Status

(Figure 3) shows the existence of a peak in the evolution of the number of subscribers during the year 2012. The reasons for this rise can be explained by the massive agreement of loans ANSEJ, CNAC, etc., and by the Algerian government decisions during this year.

Indeed, to be eligible for the loan, the bank requires an insurance contract from the owners generally it is MP contracts.

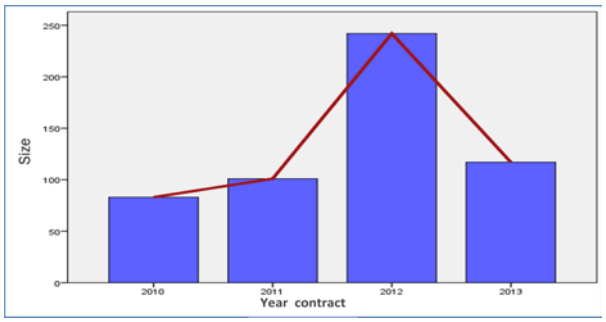


Figure 3. The number of subscribers per year

Following this, the SAA [6] has established agreements with these organizations to expand and maintain its leading position in the Algerian market. On the other side, and in order to optimize the customer portfolio of SAA in simple risk, the company has adopted a policy of encouragement by granting reductions on the automobile contract.

But this reduction can't be obtained without a subscription to a MH contract, CAT-NAT.

In (Figure 4), it is observed that the most frequent incident is the DDE with a percentage of 86.6%. These results are in agreement with reality. Indeed, most of the disasters reported are of the DDE type. This is due to various structural and cultural factors (bad quality piping, maintenance negligence, irresponsible citizens).

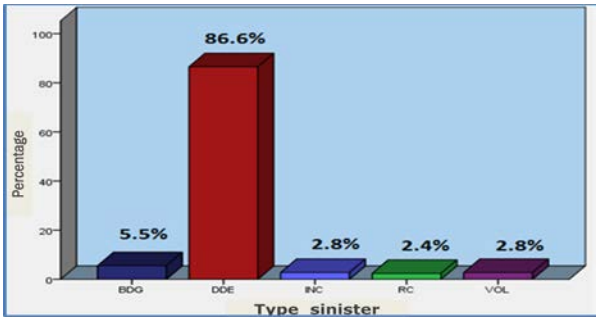


Figure 4. Distribution of insured by Type of Disaster

3. Extraction of Knowledge

In this section we follow the modeling approach of (Figure 5). We first apply binary logistic regression to evaluate the contribution of the explanatory variables to the explanation of the target variable.

We then proceed to the selection of the most relevant variables using the STEPDISC procedure based on two approaches (Ascending and Descending), and finally, we extract the characteristics of each branch by the V-test tool.

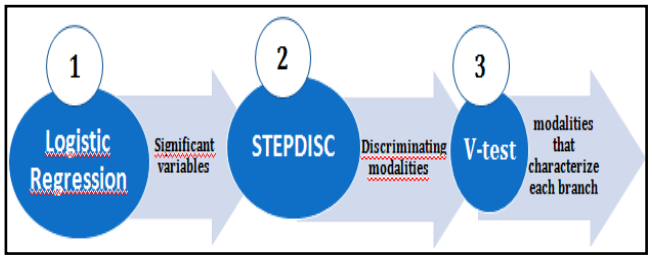


Figure 5. Modeling approach

3.1. Binary Logistic Regression

The principle of binary logistic regression [7] is based on the modeling of a variable to explain binary Y from p explanatory variables X = (x_1, x_2... x_p). The probability of belonging to an individual I has one of the two modalities of Y equal to: P (I) = P (Y = y_k / X); K = 1.2.

The logit function can be defined by the following formula:

$$\text{Ln}\left(\frac{\Pi(\omega)}{1-\Pi(\omega)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots\ldots + \beta_n x_n$$

Practically, this model estimates the coefficients β_0, β_1 … β_n. which indicate the direction and the intensity of the link of an explanatory variable on the membership of one of the branches MP or MH.

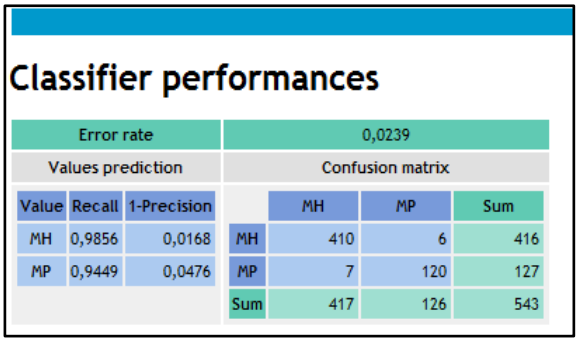
The R2 expresses the way in which the explanatory variables X contribute to the explanation of the branch variable but remains an insufficient criterion of selection.

Additional indicators are used to judge the significance of the explanatory variables (see below). Like any supervised training technique, we constituted our two samples. A learning sample (70% of observations) to learn the model and a test sample (30% of observations). The evaluation of the quality of a model is expressed by a confusion matrix.

3.1.1 Confusion Matrix

We report below (Table 2) the commented results of logistic regression.

Table 2. Confusion Matrix



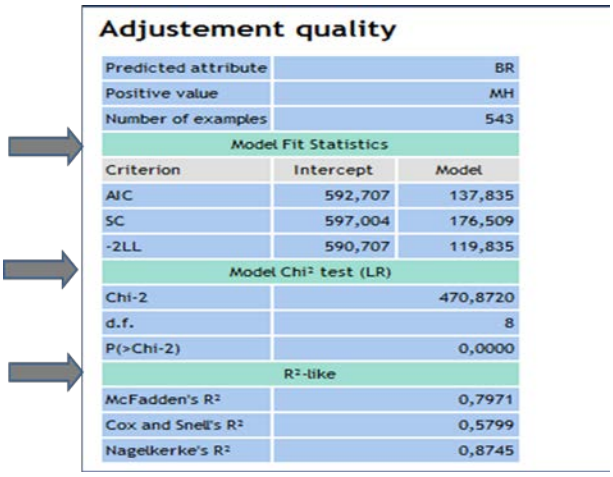
Classifier performances

Error rate			0,0239		
Values prediction			Confusion matrix		
Value	Recall	1-Precision		MH	MP	Sum
MH	0,9856	0,0168	MH	410	6	416
MP	0,9449	0,0476	MP	7	120	127
			Sum	417	126	543

From this matrix, it can be seen that the classification error rate of our model is equal to 0.0239. The model has a good ranking rate, estimated at 97%, which allows us to confirm the good quality of the explanatory variables.

3.1.2. Evaluation of the Model

Table 3. Evaluation of the Model



Adjustement quality

Predicted attribute		BR
Positive value		MH
Number of examples		543
Model Fit Statistics		
Criterion	Intercept	Model
AIC	592,707	137,835
SC	597,004	176,509
-2LL	590,707	119,835
Model Chi² test (LR)		
Chi-2		470,8720
d.f.		8
P(>Chi-2)		0,0000
R²-like		
McFadden's R²		0,7971
Cox and Snell's R²		0,5799
Nagelkerke's R²		0,8745

The section: "**Model fit statistics**" of (Table3) compares the studied model with the INTERCEPT trivial model "reduced to constant".

The best model is the one for which AIC [8] and SC are the lowest. According to the results, it can be concluded that the studied model is better. In addition, the chi-square statistic at the 5% threshold is globally significant.

The R2-like (Pseudo-R^2) is near to 1, so the regression model is good. This leads us to say that the explanatory variables contribute to the explanation of the "branch" variable.

3.1.3. Significance of Parameters

As shown in (Table4), the most discriminating variables are: *legal status, state, insurance premium, evaluation and amount of the sinister.*

Attributes in the equation

Attribute	Coef.	Wald	Signif
constant	57,133572	0,0098	0,9211
legal_status	-3,810069	10,3427	0,0013
Type_sinst	0,588101	3,1925	0,0740
State	-0,930823	4,0234	0,0449
year_contrat	-0,822362	1,7937	0,1805
year_disaster	0,801800	2,0960	0,1477
premium	-2,565350	23,1239	0,0000
Eval_sinst	-6,194058	65,1609	0,0000
Amount_sinst	0,976541	4,7155	0,0299

Table 4. Estimation of Parameters

3.1.4. Comment

We can conclude that the model is good and effective on our data file; the explanatory variables contribute significantly to the explanation of the target variable, which made it possible to identify the most significant attributes. This first evaluation by binary logistic regression already gives us an overview of the results, which we will consolidate by other methods.

3.2. The TANAGRA StepDisc Procedure

It is important to notice that the main objective of our study is to extract the characteristics[2] of each branch (MH and MP). In order to do that, we consider all the modalities of the variables in the model, which gives us a summary of 28 variables.

To reduce this large number of variables, we have chosen the STEPDISC method [9] [10] which allow selecting step by step the variables potentially interesting. To give more credibility to our results, we apply the two strategies of the procedure: FORWARD and BACKWARD, by choosing Fisher's statistic as a stop rule at the 3.84 threshold.

The variables that have more than the significance threshold are given in the table below.

Thus, of the 28 variables, the procedure selected only 10 (see Figure5).

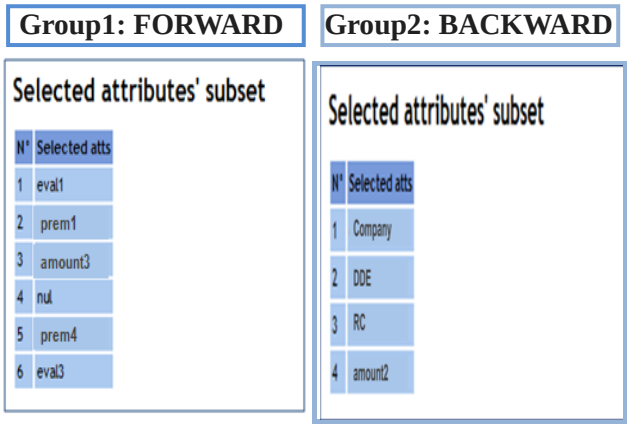


Figure 5. Variables Extracted by STEPDISC

3.3. The V-test

This concept, widely used in data mining, is present in the range of tools offered by most of Data Mining software. It is a criterion of predilection; it is applicable to large tables. The test-value (or V-test) performs a scheduling on the attributes. The Attributes whose v-test^2 is greater than 2 in absolute value are those which best characterize a given group.

We apply the V-test [11] [12] to the set of modalities used for each of the groups in the Branch class. The summary of the results obtained by the V-test is presented in (Figure 6).

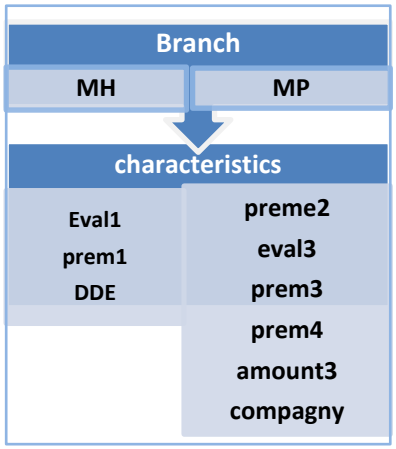


Figure 6. Characteristics of each Branch, Obtained by the V-test

3.4. Discussion

We have seen that for the MH branch, the evaluation of the amount of the sinister is lower than the MP branch. This result was predictable because the MH branch concerns much more individuals whose damage are often less severe than for businesses. This explains the high amounts of evaluation for the MP branch. It can be concluded in the same way regarding the insurance premium and the amount of the sinister.

4. Conclusion

We can observe today, that more and more applications need to integrate an exploratory analysis in order to select the useful variables to consider, for example to make a segmentation of clientele linked to the pricing. In practice, business experts, develop their own applications using conventional statistics, but in their analysis slips a great deal of subjectivity. This new methodology, which we have applied in this paper, is not very widespread because the introduction of data mining methods [13] in the field of insurance is relatively recent.

That could potentially lead to the idea of the personalized tariffing involving a paradox with the basic principle of mutualisation of the risks at the origin of insurances.

In this paper, we got into an important problem which is that of the discovery of the relevant variables, using an innovative approach based on the data mining methods. In a supervised learning framework, the modeling process adopted aims at finding the characteristics by linking the variable to explain the branch to the other explanatory variables, in order to identify all the full set of explanatory variables likely to interest the insurer. These variables are classified by decreasing order by the V-test.

The approach that we applied has allowed to bring out the explanatory variables most discriminating the according to the branch. Thus, The MH branch is characterized by the occurrence of water damage, value of evaluation of the disaster, and a minimal insurance premium. On the other hand, the MP Branch is identified by the amounts of insurance premium, the evaluation of the sinister and the highest amounts.

We could observe that the characteristics of the products MH and MP are distinct. Each product has the characteristics which are specific to it. These results will allow insurers to adapt the offerings of each branch to the type of accident and their severity and the legal status is to the needs of the insured. The objective is to arrive at the idea of customized pricing for insurance premiums, resulting in a paradox with the basic principle of mutualisation of risks at the origin of insurance[3].

[3] http://www.br3consultants.fr/2016/03/17/les-nouveaux-enjeux-de-lassurance-entre-segmentation-et-mutualisation/

In conclusion, it is important to say that *the suggested approach is generic and can be applied to other domains.*

Acknowledgements

I would like to thank Yasmine for their help.

References

[1]. F. Noël. La gestion des sinistres IRD incendies et risques divers, édition séfi, Canada 2014.

[2]. S. Tuffery. Data mining et statistique décisionnelle 4ième ed, Broché – 21 août 2012.

[3]. http://www.economie.gouv.fr/dgccrf/Publications/Vie-pratique/Fiches-pratiques/Assurance-multirisque-habitation 2014-04-22.

[4]. M. Boullé. Recherche d'une représentation des données efficace pour la fouille des grandes bases de données. Ph. D. thesis, ENST. 2007.

[5]. D. Pyle. Data preparation for data mining. Morgan Kaufmann Publishers, Inc. San Francisco, USA. 1999

[6]. J. M. Hilbe. Logistic regression models. Chapman & Hall / CRC Texts in Statistical Science Series. CRC Press, Boca Ra-ton, FL,2009. ISBN978-1-4200-7575-5.

[7]. H. Akaike. A new look at statistical model identification. IEEE Transactions on Automatic Control AU-19: 716-722. 1974

[8]. M. Bardos Analyse discriminante « application au risque et scoring financier » édition Dunod, PARIS 2011

[9]. http://eric.univ-lyon2.fr/_ricco/tanagra/en/tanagra/

[10]. L. Lebart & A. Morineau & M. Piron."Statistique Exploratoire Multidi-mensionnelle" Dunod, pp.181-184, 2000.

[11]. L. Lebart, A. Morineau, M. Piron (1997). Statistique Exploratoire Multidimensionnelle. Dunod.

[12]. C. Wesphal & T. Blaxton. "Data Mining Solutions", John Wiley, New York, 1998.

Published: Volume 2017, Issue 11 / November 25, 2017