Setting up a Mechanism for Predicting Automobile Customer Defection at SAHAM Insurance (Cameroon)

Rhode Ghislaine Nguewo Ngassam[1], Jean Robert Kala Kamdjoug[1](✉), and Samuel Fosso Wamba[2,3]

[1] Université Catholique d'Afrique Centrale, FSSG, GRIAGES, Yaounde, Cameroun
ghislainengassam5@gmail.com, jrkala@gmail.com
[2] Toulouse Business School, Toulouse, France
s.fosso-wamba@tbs-education.fr
[3] Université Fédérale de Toulouse Midi-Pyrénées, 20 Boulevard Lascrosses,
31068 Toulouse, France

Abstract. As markets become more competitive, companies have realized the need to manage the loss of customers (Churn) especially in terms of its prediction. To achieve this, in datamining framework, the main challenge is the selection of variables and the technique adapted to the studied context. This article examines the case of SAHAM insurance and uses ANOVA, chi-square test and Pearson correlations table for variable selection. To make an objective decision on selection of a technique among others, the multi criteria decision aid method PROMETHEE-GAIA has been used. With the aim to improve the initial model, which results was mitigated; the data set has been separated in two groups: individual customers and corporations. Then, with computation of the new one, we observe that, in general, performance is better on the group of individual customers than on previous global model and on corporations.

Keywords: Churn · Data mining · Decision tree · PROMETHEE GAIA
ANOVA · Chi square · Pearson correlations

1 Introduction

The saturation of today's markets is pushing companies to develop new strategies to capture and retain customers. Several studies show that it is better to retain existing clients than lure new ones [1–3], and that managing the loss of customers, otherwise known as Churn, is of paramount importance.

The Churn is the combination of two terms "Change and Turn" [4] and signifies the loss of customers in view of competition [5]. Interest in this topic resides in the fact that loyalty policies need to be targeted at customers with a high risk of churning [6], and this should not be done globally to avoid wasting resources and inefficiency of such policies [7]. This explains why numerous studies have been carried out on this subject in relation to other fields, such as telecommunications [8], banking [7], press subscriptions [1], conventional and internet shopping [9], and even in the insurance sector [10, 11].

Á. Rocha et al. (Eds.): WorldCIST'18 2018, AISC 746, pp. 878–888, 2018.
https://doi.org/10.1007/978-3-319-77712-2_83

In Cameroon, competition has become fiercer than ever the insurance market, with a total of fourteen (14) companies delivering mainly automobile insurance policies [12–14]. As a result, the customer portfolios of different insurance companies are being constantly weakened because a customer can at any time decide to abandon their traditional insurer for another operator. In the case of SAHAM insurance company, the problem is very prominent as its Churn rate varied between 55% and 61% in 2012–2016.

The problem of Churn mastery is very often summed up in its prediction or detection [15]. This involves the search for customer data through data mining techniques, so as to set up a powerful customer detection model that can help identify those who are likely to "Churn" [9, 10, 16, 17]. In this regard, it is necessary to identify the relevant, best suited variables and technique for each context.

This paper uses the case of SAHAM Insurance (Cameroon) to construct a predictive model for Churn. The following sections present a literature review on Churn, the research methodology that we have adopted for this purpose, the results obtained and their discussion, as well the research implications and the conclusion.

2 Literature Review

Generally, the contextual prediction of Churn can be comparable to a knowledge discovery process present by Alsultanny [18]. The following chart summarizes this approach (Fig. 1).

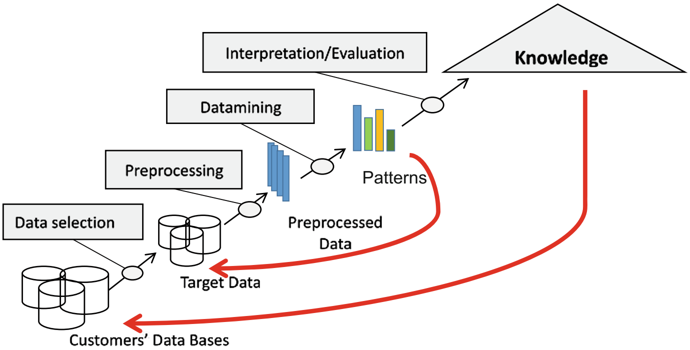

Fig. 1. Knowledge discovery process

2.1 Data Selection and Preprocessing

The problematic of data selection can be comparable to the selection of variables in the sense that the challenge is often to choose relevant variables between those found in the literature and between many others that can be extracted in customer's databases [1, 19, 20].

Some authors use variables identified in the literature and those suggested to them by the experts of the field and eliminate those of these variables which are not available in customer's databases of the considered company [11]. This approach is very simplistic and does not take into account the fact that all the variables selected may not be relevant enough for the context under study in that they may not carry enough predictive information.

To overcome the limits of this first approach, there are in the literature very specific techniques for eliminating objectively irrelevant variables without risk of harming the performance of the models to be built or prevent the results not to be generalizable [1]. In this sense, Farquad et al. [21], in his paper related to Churn prediction using SMV (Support Vector Machine), used the SMV-RFE (SMV-Recursive Feature Elimination) algorithm to select the variables of his study realized in a Latin American bank. In the same way, Tsai and Chen [22], in their paper on the selection of variables, used the association rules to select the variables to use in the construction of his model in a company of multimedia on demand and [23] uses the AUC parameter-selection technique in B2B commerce industry.

From among the many techniques that can be used to identify and select these variables, our study has resorted a two-step process consisting in (i) selecting the variables that significantly influence the target variable [1, 24], and (ii) suppressing any redundancy [1], i.e. removing those variables that are highly correlated with others.

After selecting variables, we have to extract and prepare the dataset by deleting non-available values and reflecting data changes over time [7].

2.2 Data Mining

Data mining techniques are generally used for the extraction of unknown information that can be used for decision support in companies. However, the challenges is often to know which technique to use for a specific study as there is a large number of techniques in the literature such as decision tree, logistic regression, neural network, support vector machine or association rules [22, 25] if we can only mention those ones.

The choice of the technique to use in a study is not always a very meticulous process as demonstrated in the extant literature, as the authors generally prefer the criterion of popularity of techniques [26, 27], and proceed by comparing the models resulting from the different techniques being used [11, 25], or they simply choose one technique and make a demonstration of its relevance and applicability [1, 21, 27].

Another authors have adopted a different approach to select one technique by describing the studied environment before looking for a technique to match with [22]. In the same way, [28] has resorted a table in which many techniques are compared from the point of view of several characteristics.

However, each technique has its own characteristics and advantages, which the authors generally describe based on three main criteria: (1) performance, which depends on the characteristics of the dataset to be studied [26, 28]; (2) interpretation, which measures the ability for a technique to give results that are directly usable [21]; and (3) the processing time, which takes into account the training time and other tasks to achieve usable results [10, 28].

2.3 Interpretation and Evaluation of the Model

Following the selection of variables and techniques, a model witch can be presented as equation for parametric technique and graphically for non- parametric technique but the aim is always to extract decision rules [21]. When the model is constructed, its performance can be measured using coincidence matrices [27, 29, 30] (Table 1).

Table 1. Coincidence matrices.

	Churn	Fidelity
Churn	A	B
Fidelity	C	D

A refers to the number of customers who are predicted to “churn” and who “churn”. C refers to the number of customers who are predicted to “churn” but who do not “churn”. B is the number of customers who are predicted not to “churn” but who “churn”. D refers to the number of customers who are predicted not to “churn” and who do not “churn”.

From the matrices, we can calculate the following performance indicators: recall rate (R), precision rate (P) and F-measure (F) [27, 29].

$$R = \frac{A}{A+B}; \quad P = \frac{A}{A+C}; \quad F = \frac{2 * R * P}{R+P}$$

3 Methodology

We adopted a methodology in three main stages: the selection of variables, the selection of a data mining technique and the phase of data extraction and model construction.

3.1 Variable Selection

The choice of variables to be considered during the study is a statistical challenge in which we seek to retain the most relevant variables while eliminating duplication. We will use the ANOVA test, the Chi-square test and the Pearson correlations table. The Table 2 below presents variables identified in both the literature and the SAHAM insurance company’s database for automobile policy.

The ANOVA test is carried out to check whether by using the Fisher’s test, the distribution of each of the variables is related to the modalities for target variables [24].

The ANOVA test starts with the following two working hypotheses: (1) The experimental unit is confused with the statistical individual; and (2) The plan is completely randomized and the modalities for the factor are randomly assigned to the experimental units.

Table 2. Description of variables and origins.

Variable	Type	References
ID	Text	Adapted from [11]
Category	Qualitative	Adapted from [1]
Contract duration	Quantitative	Adapted from [22]
Number of subscriptions, 2012	Quantitative	Interviews with managers
Number of subscriptions, 2013	Quantitative	
Number of subscriptions, 2014	Quantitative	
Number of subscriptions, 2015	Quantitative	
Number of prior defections	Quantitative	
Premium, 2015	Quantitative	Adapted from [6]
Premium, 2014	Quantitative	
Premium 2013	Quantitative	
Premium, 2012	Quantitative	
Number of losses, 2012	Quantitative	Adapted from [11]
Number of losses, 2013	Quantitative	
Number of losses, 2014	Quantitative	
Number of losses, 2015	Quantitative	
Compensation period, 2012	Quantitative	Adapted from interviews with managers
Compensation period, 2013	Quantitative	
Compensation period, 2014	Quantitative	
Compensation period, 2015	Quantitative	
Subscription period, 2012	Quantitative	Adapted from [1]
Subscription period, 2013	Quantitative	
Subscription period, 2014	Quantitative	
Subscription period, 2015	Quantitative	
Status	Qualitative	Is the target variable

In our context, we have a population P and a class B for our study (Churner, Faithful). The population P will be divided into two subpopulations Pc and Pf, respectively with the means μc and μf for the variable X, the global average of this variable being μ. The ANOVA test is based on the following hypothesis test:

$$\begin{cases} H_0{:}\mu_c = \mu_f = \mu \\ H_1{:}\mu_c \neq \mu_f \end{cases}$$

With regard to the qualitative variables, we will test their independence in relation to the target variable from a Chi-square test. This test can be likened to a collinearity test as they are related to quantitative variables. Therefore, we set out the following two hypotheses:

H0: The categories of customers and status variables are independent of each other.
H1: The categories of customers and status variables are statistically interdependent.

The Pearson correlations table lists all correlation coefficients between variables. In this table, we can eliminate those variables that are highly correlated.

3.2 Selection of the Relevant Technique

The selection techniques that are frequently used in previous research works include the decision trees, the logistic regression, the neural networks and the machine support vectors. As mentioned above, the authors describe data mining techniques according to three main criteria. A comparative table can be obtained in this study [28].

Such technique has been selected considering the multi-criteria nature of decision support. While it is influenced by such criteria, especially the performance criterion, as well as the characteristics of the data, it is also necessary to consider the characteristics of service end-users and the expectations of decision-makers. Therefore, the choice of one technique compared to another with regard to the criteria is relative, as there is not a better method in relation to another for all the criteria, so using an objective method for the choice process is critical [31].

As a result, we have chosen to work with the PROMETHEE GAIA method, which is one of the latest developments in decision support methods [31]. This method (PROMETHEE-GAIA -Preference Ranking Organization Method for Enrichment Evaluations-Geometrical Analysis for Interactive Aid) is a prescriptive approach to multi criteria problem analysis with several actions (or decisions) evaluated according to several criteria. Such an approach enables actors to visualize conflicts and synergies between criteria during the decision process. Using the Visual PROMETHEE software, we performed not only the analysis with the same weight for the criteria, but also the qualitative evaluation of five scales. Then, we varied the weights based on discussions and interviews with the managers of SAHAM (e.g., the Marketing Director and the Director of Audit and Quality), using the “walking weight” module in Visual PROMOTHEE.

3.3 Data Extraction and Model

We extracted the data using the selected variables from the business software (ORAS 7/MILLIARD) from the company SAHAM assurance. The dataset consists of 3 190 “non-churners”, 2 621 “churners”; of 5 168 individuals, 643 corporations.

With the R statistical software, we deleted the empty values and then constructed a model using the technique that was previously selected. The training step was done on 75% of the data set and the test phase on the rest. The next step consisted in evaluating the model with the coincidence matrices, followed by a study of the behavior of the performance variable as we altered the dataset.

4 Results and Discussion

The results can be presented following the methodology approach on three steps.

4.1 Variables Selected

With regard to the available variables, we first selected those in the database before embarking on the ANOVA test, the chi-square test, and the Pearson correlations.

The chi-square test revealed a significant dependence between the various categories of customers and its statuses and the Pearson correlations table shows that there is none high correlation between variables in the Table 3 and the category. So we will keep all these variables for the rest of our work.

Table 3. ANOVA Test result of the whole dataset - significant link.

Variables	F-value	Pr (>F)	P-value
Contract duration	71.26	2.00E−16	0.000***
Number of subscriptions, 2015	192.7	2.00E−16	0.000***
Number of subscriptions, 2012	102.6	2.00E−16	0.000***
Number of subscriptions, 2013	85.65	2.00E−16	0.000***
Number of subscriptions, 2014	103.1	2.00E−16	0.000***
Number of prior defections	286.3	2.00E−16	0.000***
Subscription period, 2014	7.398	0.00655	0.001**
Premium, 2015	6.568	0.0104	0.05*
Subscription period, 2015	5.96	0.0147	0.05*

4.2 Technique Selected

The selection process of the method was done together with SAHAM managers (Marketing Director, Director of Audit and Quality) using Visual PROMETHEE in three steps. (1) Firstly, we described the methods using criteria which had the same weight at the beginning in the Visual PROMETHEE software. The primary result was to be discussed. (2) Secondly, considering the “walking weight” module, the decision-makers suggested some adjustments of the criterion weights in an interactive process where the model was computed. The results were presented to decision-makers and adjustments were made. Lastly, the final results guiding the choice of method were validated as we can see in Fig. 2 below.

The Decision tree is the most suitable method according to SAHAM’s senior management and the beneficiaries of the project, who indicate six advantages of the method: (1) it can be equally effective for both quantitative and qualitative variables; (2) the nonlinear relations between variables are considered; (3) the method is effective on large samples; (4) the method gives easily interpretable results; (5) and it has a short training time.

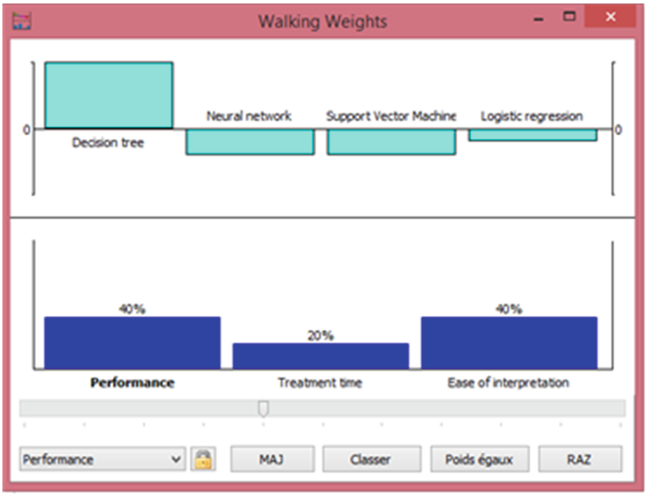

Fig. 2. PROMETHEE recommendation.

4.3 Construction and Evaluation of the Model

The constructed decision tree consists of three nodes and four leaves. We can break down this tree into four main decision rules and the variables of the three nodes are: the premium 2015, the number of defections and the duration of the contract. The model performed as follows (Table 4):

Table 4. Coincidence matrices for the general model – test sample.

Actual/model	Churn	Not-churn
Churn	377	285
Non-churn	388	694

This table allows us to calculate the following indicators in the Table 5 below.

Table 5. Performance indicators for the general model – test sample.

	Precision	Recall	F-measure
Churn	49.281%	56.95%	52.83%
Non-churn	70.888%	64.14%	67.35%

Given that the results were very average and that the category variable is linked to the target variable, we separated the two categories of customers and resumed the same process. We noticed that the remaining variables did not influence the target variable in the same way in each category i.e. the decision trees of the two categories haven't the same variables on their nodes. There is also a slight improvement in the performance of the models constructed especially based on the category of individual customers (Table 6).

Table 6. Performance indicators for the two partial models.

	Precision		Recall		F-measure	
	Churn	Non-churn	Churn	Non-churn	Churn	Non-churn
Individual	47.78%	76.64%	63.23%	63.59%	54.43%	69.51%
Corporation	35.94%	84.45%	60.53%	66.67%	45.10%	74.51%

Another observation is that by dividing the dataset, the recall rate is improved, so each group of customers has its own behavior and is not influenced by the same parameters (variables).

When we consider the dataset entirely, we have a decision tree with three variables: premium 2015, number of prior defections, and duration of the contract. However, when we take the datasets of individual customers, we have a decision tree with two mains parameters: premium 2015, and number of prior defections. With the group of corporations, we obtained a decision tree with two variables as well: duration of the contract, and number of subscription 2015.

Moreover, we observe that, in general, performance is better on the group of individual customers than on previous global model and on corporations.

5 Managerial Implications

The variables considered by the model represent parameters that the business management must follow to avoid the defection of customers. For example, customers with the premium 2015 below a certain amount will need more attention, especially if this feature is combined with other data by the built decision tree.

In addition, each group of customers has a particular understanding, so it is necessary to perform a more detailed segmentation of the market before looking for the variables that influence each group as well as the decision rules regarding the model.

6 Conclusion and Future Directions

When it comes to setting up a mechanism for predicting automobile customer defection in an insurance environment, this paper has demonstrated how to go about the selection of the variables and techniques, and how the customer department can be divided into homogeneous groups for increased prediction performance.

Through various techniques such as the ANOVA, chi square test and Pearson correlations, and by using PROMETHEE GAIA (a method for multi-criteria decision support), we were able to select from a set of variables those that were significantly related to the target variable, but also we could see that the decision tree was the most appropriate technique for the database of the SAHAM Insurance Company (such data were both quantitative and qualitative, with non-linear relationships between certain variables, with extreme values and with a larger or smaller working sample), and considering the characteristics of the organization (which favors quick results and ease of interpretation, given the profile of the beneficiaries).

This study has a theoretical contribution to the prediction of Churn in the automobile insurance industry in Cameroon and can be used in similar studies. However, future studies could focus on conducting a more detailed study of other useful methods that can give rise to more relevant criteria for the selection of the right technique according to each context. Another angle of research could be directed at the impact of variations in extreme values on model performance.

References

1. Coussement, K., Benoit, D.F., Van den Poel, D.: Improved marketing decision making in a customer churn prediction context using generalized additive models. Expert Syst. Appl. **37**(3), 2132–2143 (2010)
2. Reichheld, F.F., Sasser, W.E.: Zero defections: quality comes to services. Harvard Bus. Rev. **68**, 105–111 (1990)
3. Gremler, D.D., Brown, S.W.: The loyalty ripple effect: appreciating the full value of customers. Int. J. Serv. Ind. Manag. **10**(3), 271–293 (1999)
4. Yabas, U., Cankaya, H.C.: Churn prediction in subscriber management for mobile and wireless communications services. In: 2013 IEEE Globecom Workshops (GC Wkshps). IEEE (2013)
5. Zhao, Y., et al.: Customer churn prediction using improved one-class support vector machine. In: International Conference on Advanced Data Mining and Applications. Springer (2005)
6. Hadden, J., et al.: Computer assisted customer churn management: state-of-the-art and future trends. Comput. Oper. Res. **34**(10), 2902–2917 (2007)
7. Hu, X.: A data mining approach for retailing bank customer attrition analysis. Appl. Intell. **22**(1), 47–60 (2005)
8. Huang, B., Kechadi, M.T., Buckley, B.: Customer churn prediction in telecommunications. Expert Syst. Appl. **39**(1), 1414–1425 (2012)
9. Song, H.S., Kim, J.K., Kim, S.H.: Mining the change of customer behavior in an internet shopping mall. Expert Syst. Appl. **21**(3), 157–168 (2001)
10. Pannetier Lebeuf, S.: Prédiction de l'attrition en date de renouvellement en assurance automobile avec processus gaussiens (2011)
11. Huigevoort, C.W.J.M.: Customer churn prediction for an insurance company, Master of Science: Information System, p. 99. Eindhoven University of Technology, Eindhoven (2015)
12. ASAC: Assurance et Sécurité: le marché camerounais de l'assurance. Magazine de L'ASAC, No. 39, pp. 18–34 (2017)
13. ASAC: Assurance et Sécurité. Magazine de l'Association des Sociétés d'assurances au Cameroun, 2016. No. 38, Décembre 2016
14. ASAC: Rapport sur le marché camerounais des assurances: Exercice 2015. Statistiques, pp. 7–8 (2015)
15. Neslin, S.A., et al.: Defection detection: measuring and understanding the predictive accuracy of customer churn models. J. Mark. Res. **43**(2), 204–211 (2006)
16. Ngai, E.W., Xiu, L., Chau, D.C.: Application of data mining techniques in customer relationship management: a literature review and classification. Expert Syst. Appl. **36**(2), 2592–2602 (2009)
17. Wei, C.-P., Chiu, I.-T.: Turning telecommunications call details to churn prediction: a data mining approach. Expert Syst. Appl. **23**, 103–112 (2002)
18. Alsultanny, Y.A.: Database preprocessing and comparison between data mining methods. Int. J. New Comput. Archit. Appl. (IJNCAA) **1**(1), 61–73 (2011)

19. Blum, A.L., Langley, P.: Selection of relevant features and examples in machine learning. Artif. Intell. **97**(1), 245–271 (1997)
20. Dash, M., Liu, H.: Feature selection for classification. Intell. Data Anal. **1**(1–4), 131–156 (1997)
21. Farquad, M.A.H., Ravi, V., Raju, S.B.: Churn prediction using comprehensible support vector machine: an analytical CRM application. Appl. Soft Comput. **19**, 31–40 (2014)
22. Tsai, C.-F., Chen, M.-Y.: Variable selection by association rules for customer churn prediction of multimedia on demand. Expert Syst. Appl. **37**(3), 2006–2015 (2010)
23. Gordini, N., Veglio, V.: Architectures: customers churn prediction and marketing retention strategies. An application of support vector machines based on the AUC parameter-selection technique in B2B e-commerce industry. Ind. Mark. Manag. **62**(Supplement C), 100–107 (2017)
24. Baccini, A., et al.: Pour l'analyse statistique de données transcriptomiques. Journal de la société française de statistique **146**(1–2), 5–44 (2005)
25. Wu, H.-L., Zhang, W.-W., Zhang, Y.-Y.: An empirical study of customer churn in e-commerce based on data mining. In: 2010 International Conference on Management and Service Science (MASS). IEEE (2010)
26. Ali, Ö.G., Arıtürk, U.: Dynamic churn prediction framework with more effective use of rare event data: the case of private banking. Expert Syst. Appl. **41**(17), 7889–7903 (2014)
27. He, Y., He, Z., Zhang, D.: A study on prediction of customer churn in fixed communication network based on data mining. In: Sixth International Conference on Fuzzy Systems and Knowledge Discovery, FSKD 2009. IEEE (2009)
28. Risselada, H., Verhoef, P.C., Bijmolt, T.H.: Staying power of churn prediction models. J. Interact. Mark. **24**(3), 198–208 (2010)
29. Bin, L., Peiji, S., Juan, L.: Customer churn prediction based on the decision tree in personal handyphone system service. In: 2007 International Conference on Service Systems and Service Management. IEEE (2007)
30. Chawla, N.V.: Data mining for imbalanced datasets: an overview. In: Data Mining and Knowledge Discovery Handbook, pp. 875–886. Springer (2009)
31. Brans, J.-P., Mareschal, B.: PROMETHEE methods. In: Multiple Criteria Decision Analysis: State of the Art Surveys, pp. 163–186 (2005)