DOI: 10.2507/31st.daaam.proceedings.112

MACHINE LEARNING ALGORITHMS IN THE PROFITABILITY ANALYSIS OF CASCO INSURANCE

Davor Skobic, Goran Kraljevic & Marin Mandic

This Publication has to be referred as: Skobic, D[avor]; Kraljevic, G[oran] & Mandic, M[arin] (2020). Machine Learning Algorithms in the Profitability Analysis of Casco Insurance, Proceedings of the 31st DAAAM International Symposium, pp.0807-0811, B. Katalinic (Ed.), Published by DAAAM International, ISBN 978-3-902734-29-7, ISSN 1726-9679, Vienna, Austria
DOI: 10.2507/31st.daaam.proceedings.112

Abstract

Main aim of this work is on creating predictive model which would enable observation of behaviour patterns of the users of casco car insurance and predict their profitability in the future. Creation of a good model depends on analysis of available data and its preparation. Data preparation is defined as acknowledgment and retrievement of relevant data, cleaning and unifying records. There have been created and compared multiple models based on different methods and algorithms of machine learning (decision tree, neural networks, logistic regression). For modelling examples RapidMiner open-source analytical tool has been used.

Keywords: Predictive model; Profitability analysis; Data mining algorithms.

1. Introduction

In the insurance industry main goal is to cover unexpected events to provide clients with safety, as well as to make profit. Our goal for this paper is to make data mining model for user profitability trend prediction. Recent research of data mining usage in insurance goes in multiple directions. In paper [1] we can see that Customer Relationship Management is an area where data mining is used so insurance can realise true value of a customer. In paper [2] we can see a data mining method called decision tree used for client classification in car insurance by their preferential group. Also, as said by paper [3] we can do client character judgement if we possess quality data. In papers [4] and [5] there has been shown usage of data mining in risk management for the insurance.

As said in paper [6] it is considered that 20 to 25 per cent of claims contains some kind of fraud level, which results in 10 per cent increase in claim payments. In papers [7] and [8] we can see usage of KDD, data mining, machine learning and statistics for fraud detection. Also, data mining has been used in other industries such as manufacturing [9], finance [10] and telecommunications [11]. This paper covers profitability analysis of clients in casco insurance, and application of this research can be found in fields of risk management and customer relationship management. This research included the following steps: data preparation, modelling and result analysis. Each of these steps will be described in following chapters. The conclusion of this paper goes in the direction of checking whether it is possible to predict the profitability of casco insurance users.

2. Data preparation

For quality data analysis, understanding of its business scope is necessary. Client profitability analysis in this paper is being done on every insurance policy concluded among the client and insurance company in question.

Each row of our data is part of one of these categories:

- Client,
- Vehicle,
- Policy data,
- Business data.

For good model creation it is necessary to do available data analysis and its preparation. Relevant data must be queried, cleaned and unified. Also, rows with missing data must be filled out or removed. Data on which this paper is based on is queried from databases of an insurance company. Data has been cleaned by ETL procedures, since data has to be formatted and value unified in data warehouse. Data mining tool RapidMiner has functionality for these procedures. In the following table used attributes of data for data mining are visible. On certain attributes there has been discretization performed for the purpose of easier data mining process.

Category	Attribute	Value range
Client	Age	1. 18-30 2. 31-45 3. 46-60 4. >60
Policy	Previous policy	True/False
Client	Type	1. Natural person 2. Legal entity
Client	Gender	1. Male 2. Female 3. Legal entity
Client	Life insurance	True/False
Policy	Organizational unit	1 – 5
Policy	Duration	1. One year 2. Less than a year 3. More than a year
Policy	Number of instalments	One/More
Policy	Premium	1 – 4
Policy	Leasing	True/False
Vehicle	Age	1 – 4
Vehicle	Brand	Multiple brands
Vehicle	Power	8 categories
Vehicle	Car type	Multiple categories
Vehicle	Price – when it was new	Multiple categories
Policy	Discount	Value of discount
Business	Casco claims through history	1. One 2. Two 3. Three and more
Business	Car insurance claims through history	1. One 2. Two 3. Three and more
Business	Number of casco policies through history	Multiple values
Business	Casco profit through history	True/False
Business	Casco profit for client	True/False

Table 1. Attributes of data mining

The target variable of this model is Business – Casco profit for client, and it will be labelled inside the modelling tool. The variable is defined as the difference between the total insurance premium and the claims paid to the insured.

3. Modelling and data analysis

Data modelling has been done with a RapidMiner tool. Data model can be seen in the following figures.

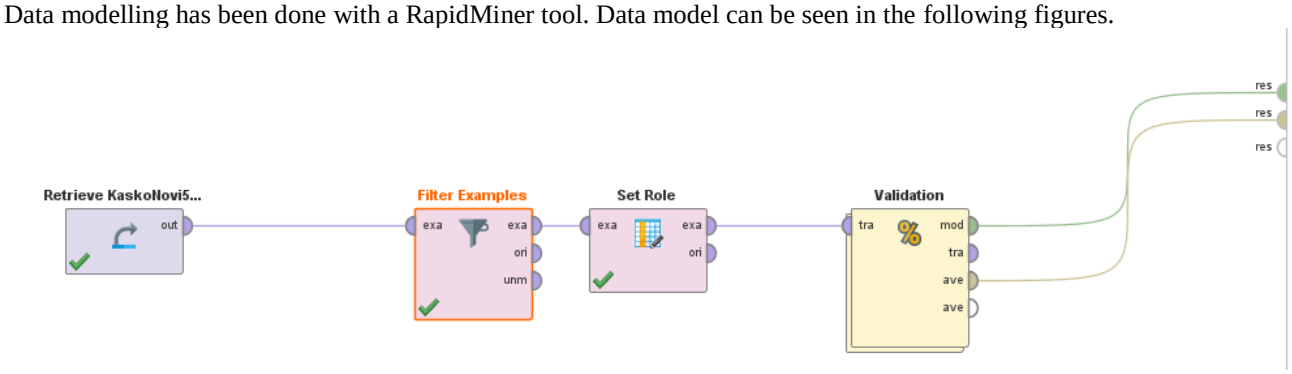

Fig. 1. Data model

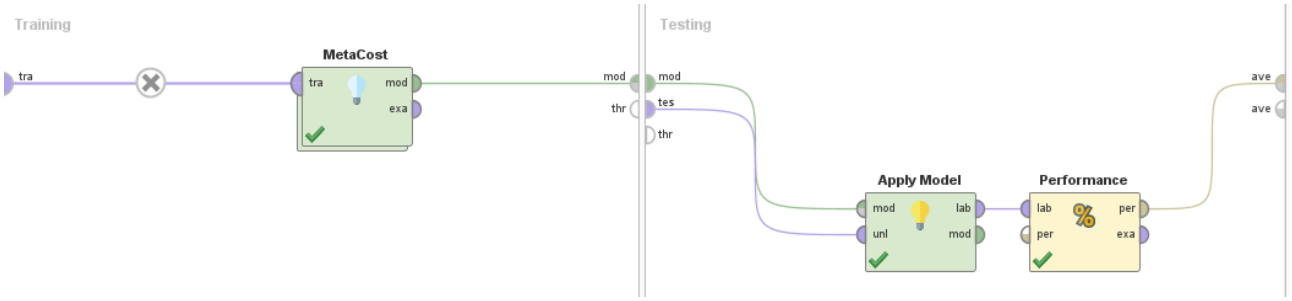

Fig. 2. Training and testing with MetaCost operator

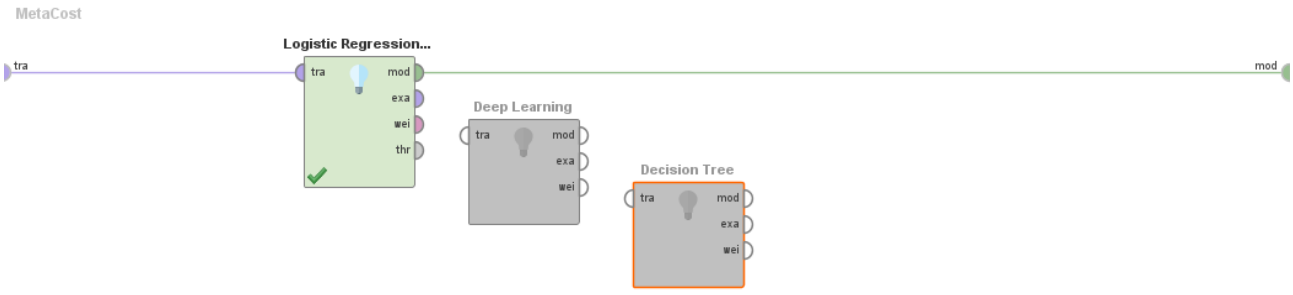

Fig. 3. Used data mining models

In Figure 1 it is visible that data has been filtered. We have used only data Client-Type-Natural person and rows without missing values. All policies that were used are from period from 2010 to 2015. Models that have been used are: Logistic regression, Decision tree and Deep learning as shown by Figure 3.

Data has been divided 70% for training and 30% for testing. The Metacost operator, as shown by Figure 2, was used to increase the cost of misclassification for non-profit clients. This metaclassifier makes its base classifier cost-sensitive by using the given cost matrix to compute label predictions according to classification costs. Results of the applied models are visible in following tables.

Cost Matrix	True Class 1	True Class 2
Predicted Class 1	0.0	1.0
Predicted Class 2	9.0	0.0

Table 2. MetaCost operator Cost matrix

	True 0	True 1	Class precision
Prediction 0	188	257	41,99%
Prediction 1	84	1339	94,10%
Class recall	68,99%	83,90%	

Table 3. Logistic regression

	True 0	True 1	Class precision
Prediction 0	175	173	50,29%
Prediction 1	95	1423	93,74%
Class recall	64,81%	89,16%	

Table 4. Deep learning

	True 0	True 1	Class precision
Prediction 0	158	29	84,49%
Prediction 1	112	1567	93,33%
Class recall	58,52%	98,25%	

Table 5. Decision tree

Logistic regression (table 3) is a specialized form of regression that is formulated to predict and explain a binary (two-group) categorical variable rather than a metric dependent measure. The form of the logistic regression variable is similar to the multiple regression variable. Deep Learning (table 4), on the other hand, is just a type of Machine Learning, inspired by the structure of a human brain. Deep learning algorithms attempt to draw similar conclusions as humans would by continually analysing data with a given logical structure.

To achieve this, deep learning uses a multi-layered structure of algorithms called neural networks. Neural networks enable us to perform many tasks, such as clustering, classification or regression. Decision tree (table 5) builds classification or regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node has two or more branches. Leaf node represents a classification or decision. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data.

Model	Accuracy	Precision	Recall	AUC
Logistic regression	81,73	94,10	83,90	0,837
Deep learning	81,51	93,89	83,83	0,852
Decision tree	92,44	93,33	98,18	0,816

Table 6. Table of measurements

In table 6 common model measurements are visible, from the RapidMiner data mining tool. In the first iteration of the launch of all models, similar results were obtained. With the desire to achieve the best possible results for the profitability category 0, the error weight operator (MetaCost) was used. In table 2 MetaCost operator Cost matrix can be seen. Logistic regression gave the best results for this category of unprofitable clients. In terms of the overall accuracy of the model, the best result was given by the Decision Tree.

All applied methods give relatively high accuracy of result prediction. A parallel with the work of churn prediction [12] [13] can be drawn, where the accuracy of the model is similar. It is certainly necessary to work on the balance of prediction, which has been attempted in this paper using the MetaCost operator. The limitations of the work are visible in the available data itself, which presents a weak point of all prediction models. The more and better data available, the more accurate the models can be. The implementation of this model would give a different insight into the profitability of the client. It could not replace conventional client selection methods, mostly due to insufficient accuracy in forecasting unprofitable clients. However, it could certainly be used as an auxiliary tool.

4. Conclusion

The main problem of this paper is whether there is a possibility to determine the model of profitability prediction. The problem was solved by creating a prediction model using existing machine learning methods. In this way, risk management and customer relations in the insurance industry, if applied, can be improved. The result of this paper is a prediction model with different results obtained by usage of different methods of machine learning.

From this paper it can be concluded that it is possible to make a prediction model about clients casco insurance profitability. This model can be used for client risk prediction, as well as the tool for CRM, since applied data mining models showed relatively high precision. Future plans go in the direction of determining the best machine learning method to apply, as well as applying different methods of data preparation. Following research should move in the direction of another kind of prediction analysis, like churn or new client risk assessment based on data mining models on history data.

5. References

[1] Ngai E.W.T., Xiu L. & Chau D.C.K. (2009), Application of data mining techniques in customer relationship management: A literature review and classification, Expert Systems with Applications, vol. 36, p. 2592-2602

[2] Xiahou J., Xu Y., Zhang S. & Liao W. (2016), Customer Profitability Analysis of Automobile Insurance Market Based on Data Mining, The 14th International Conference on Computer Science & Education (ICCSE), p. 603-609

[3] Hui S.C., Jha G. (2000), Data mining for customer service support, Information & Management, vol 38(1), p. 1-13.

[4] Jiang L. (2016), Model of the Insurance Risk Rating based on Neural Network, International Conference on Smart City and Smart Engineering, p. 375-377, 2016

[5] Jiang L. (2016), Quantitative model of Insurance Risk Management System based on Big Data, International Conference on Smart City and Smart Engineering, p. 590-593

[6] Sheshasayee A., Thomas S.S. (2017), Implementation of Data Mining Techniques in Upcoding Fraud Detection in Monetary Domains, International Conference on Innovative Mechanisms for Industry Applications, p. 730-734

[7] Bolton R.J., Hand D.J. (2002), Statistical fraud detection: A review, Statistical science, vol. 17(3), p. 235-249

[8] Yan C., Li. Y. (2015), The Identification Algorithm and Model Construction of Automobile Insurance Fraud Based on Data Mining, Fifth International Conference on Instrumentation and Measurement, Computer, Communication and Control (IMCCC)

[9] Bjorklund S. (2010), Utilisation of Data Mining Principles in Maintenance Planning, Annals of DAAAM 2010 & Proceedings of the 21th International DAAAM Symposium, Volume 21, No.1

[10] Botunac I., Panjkota A. & Matetic M. (2019). The Importance of Time Series Data Filtering for Predicting the Direction of Stock Market Movement Using Neural Networks, Proceedings of the 30th DAAAM International Symposium

[11] Pejic Bach M., Simicevic V. & Leskovic D. (2009), Microsegmentation in Telecom Market: Data Mining Approach, Chapter 93 in DAAAM International Scientific Book 2009, pp. 951-964

[12] Spiteri M., Azzopardi G. (2018), Customer Churn Prediction for a Motor Insurance Company, Thirteenth International Conference on Digital Information Management (ICDIM)

[13] Mandic M., Kraljevic G., Boban I. (2018), Performance comparison of six Data mining models for soft churn customer prediction in Telecom, International Journal of Electrical Engineering and Computing