JEAS | Journal of Economics and Administrative Sciences
E-ISSN: 2148-1792 P- ISSN: 1302-2024
Volume 3, Supplement Issue 3
www.jeasweb.org

Application of Optimized Value Techniques to Predict Damage to Fire Insurance Customers Using Data Mining

Seyed Yahya Abtahi
Assistant professor of financial engineering, Islamic Azad University, Yazd Branch, Yazd, Iran, email: yahyaabtahi@yahoo.com
Gholam Abbas Nazari Fathabad
PhD Candidate in financial engineering, Islamic Azad University, Yazd Branch, Yazd, Iran, email: nazariftth@gmail.com
Seyed Alireza Zakipour Dezfouli
PhD Candidate in financial engineering, Islamic Azad University, Yazd Branch, Yazd, Iran, email: s.alireza.zakipour@gmail.com
Rouhollah Irvani
PhD Candidate in financial engineering, Islamic Azad University, Yazd Branch, Yazd, Iran, email: Irv.reza67@gmail.com

ABSTRACT
For insurance companies it is important to know future damage value and damage distribution. Using previous damage records can help the companies to predict the future possible damage and to estimate insurance payments according to the value at risk of the customers. The purpose of this paper is to investigate the Value at Risk (VaR) using data mining. Thus, consumers are categorized in groups based on the extent of risk to which they are exposed. By the risk criteria, a given type of insurance agreement is selected for each group. Two methods of data mining and decision tree and clustering are used to devise a model for predicting customer risk in insurance industry. Variance adjustment is employed to apply the optimized VaR. The results indicated that the VaR analysis using the VaR can estimate (99%) the funds required to cover the insured damage.
Keywords: Value at Risk (VaR), simulation, data mining, insurance, damage

Introduction

With remarkable advances in information technology and overwhelming rivalry in market competitiveness, mass production of high-quality goods cannot solely ensure survival of the businesses. Increase in customer information made marketing process (Jabbar et al., 2019) more complicated and sensitive. Paying attention to management of customer relationship (Libai et al., 2020) becomes more important every day and more budget value are allocated to that in organizations (Rostami et al., 1999). One of the issues that is of great importance in the insurance industry and a great help for its growth is identifying customers and predicting their level of risk, which this study intends to examine using risk at risk using data mining in fire insurance. How to cluster customers based on their level of risk. Then, the level of risk of each of these clusters is calculated. Now new customers are placed in one of them based on which of these clusters they are more similar to, and thus the level of risk of that customer can be predicted. For this purpose, the value at risk (VAR) method (Trung, 2020) using data mining has been used to obtain decision rules and create a model for predicting the risk of customers in the insurance industry. Also, the VAR of the family is an undesirable risk measurement criterion and is a statistical

measure that quantitatively presents the maximum portfolio loss in a given period of time (Rice et al., 2020). In other words, it determines the amount of portfolio value that is expected to be lost over a period of time with a certain probability (Silahli et al., 2019). This method is currently one of the key indicators of risk measurement that financial analysts use many times.

Portfolio risk calculation consists of different types of assets, can only be measured with this method, and this is one of the benefits of value at risk, the advantage of this method is the ease of calculation and simplicity of its concept and interpretation, so that risk expresses potential assets in insurance assets quantitatively and with a number. The calculation of VAR is done in two ways, parametric and non-parametric. The parametric method consists of two basic assumptions; these two assumptions include the normality of the asset return distribution and the linearity of the relationship between market risk factors and asset value. The nonparametric method includes two techniques of historical imitation and Monte Carlo simulation (Liyanage et al., 2017). Historical simulation is a simple method because there is no need to presuppose the probability distribution of the return on assets or financial assets. Monte Carlo simulation is another non-parametric method of calculating value at risk, which has more flexibility than other methods, due to the lack of restrictions on the normalization of the probability of return on assets or the linear relationship between market risk and asset value. In the Monte Carlo method, historical information is not used, but changes are predicted using a random process and using many simulated samples that are made by data mining method. The purpose of this study is to investigate the performance of Monte Carlo simulation technique in calculating the value at risk using data mining on the daily damage values of fire insurance in Iran Insurance. Calculating the maximum potential loss in such a way that it has the lowest error rate due to market fluctuations has always attracted the attention of analysts. The use of data mining process to generate a random sample in the Monte Carlo technique is of particular importance that the objectives of the research and the distribution of data determine the process of generating random numbers.

Meybodi and Mirfakhreddini (2010) examined the risk of investing in several automobile companies. The effect of stock value fluctuations in each company on the variable was used based on the hypothetical portfolio and a suitable test was presented to evaluate the obtained credit. The results show that the stock price in Pars-Khodro Manufacturing has the highest level and in Saipa there are the least fluctuations (Vafaee Yeganeh et al., 2014). As a result, the effect of these fluctuations on the VAR of Pars-Khodro shares is more than that of Saipa Manufacturing.

In 2010, Huang tested the Monte Carlo simulation technique (Zaroni et al., 2019) to calculate the VAR, using the Brownian method to generate a random sample, and then obtained the optimal VAR by considering the adjustment factor. Finally, it examined the performance of the estimated VAR. The results of this study show that the optimal VAR correctly estimates the capital required to cover losses with a high probability. In the present study, the Monte Carlo simulation technique has been used to calculate the VAR. Random numbers were generated using data mining during the Brownian process with 5000 repetitions. To estimate the value at risk, which estimates the capital required to cover insurance liability with a 99% probability. The optimal VAR is the adjustment coefficient, which is calculated using the feedback test of the LR accuracy criterion (Gregory et al., 2020). In the continuation of the research, after reviewing data mining, in the third part, the theoretical model and the criterion for evaluating the accuracy of risk performance are presented. In the fourth section, the data are analyzed based on the results.

Methodology

Data mining: Ever since the advent of statistics, scientists have felt the need to discover the properties of data. Using statistics and methods at the time, data properties such as dispersion and concentration were examined (Chen, 2006). Statistical methods are usually suitable when it comes to examining the effect of a small number of factors on the target, but when the number of these factors increases, these methods no longer work well and in some cases are even ineffective. For example, statistical methods are less used in the analysis of lifestyle data because these data are very large. To solve this problem, scientists decided to use high-speed computers. This led to the development of other innovative methods in addition to

statistical methods such as neural networks and genetic algorithms (Lee 2006). Data mining is the "extraction of information and knowledge and the discovery of hidden patterns from a very large database." These patterns and knowledge are typically implicit (chan 2002). Data mining can be used to perform tasks such as categorizing, forecasting, estimating, and clustering data. To do this, methods have been developed that, due to the development of computers and this science, the number and quality of these methods are increasing every day. Some of the most popular of these methods are clustering orphan algorithms, neural networks, genetic algorithms, nearest neighbors, and decision trees.

Classification and clustering techniques: Classification and clustering are common issues that have been extensively studied by statisticians and machine learning researchers (Hameed et al., 2020). It is difficult to give a precise definition of these two methods, but according to the general definition, the technique of classification and clustering, separating or placing components or objects in a number of categories that do not already exist in the clustering of these categories and during the process And are created according to the properties of objects, but exist in the classification of these categories and objects are placed in these categories based on their properties (Shafiq et al., 2020; Tan and Steinbach, 2006).

Decision tree: The decision tree is a popular method for classification, the results of which are presented in a flowchart similar to the tree structure, where each node represents a test on the characteristic value and each branch represents the output of each test (Shafiq et al., 2020); the leaves of the tree also represent the categories. Typically, the complexity of a decision tree increases as the number of attributes increases. However, in some cases it has been observed that only a small number of attributes can determine the trace to which each object belongs, and the rest of the attributes are small or ineffective (Efthymia et al., 2020; Tan and Steinbach, 2006).

In constructing decision trees, data is typically divided into two categories: 1) training data used to build the model, and 2) test data used to test and evaluate the model. The quality of educational data plays a major role in determining the quality of the decision tree. If the training of the system is increased, ie the data used to train and build the model is a large percentage of the data, we will have a situation called "model over-training", which due to the presence of abnormalities in the data, training data , Produces error (Chan 2002).

Clustering and clustering: Clustering is, in fact, an unsupervised operation; this operation is used when looking for similar sets of data, without having to anticipate the similarities. Clustering is commonly used when looking for groups of customers that are not already known, ie groups do not already exist and are created in the clustering process. For example, similarities in customer use of mobile phones can be sought in order to group customers and identify new services (Borgelt2008).

Proposed method

In this research, two clustering techniques and the movement method have been used. In the clustering method, customers are divided into clusters based on their characteristics, and then the average level of damage in each of these clusters is calculated. Now the future customers, depending on which of these clusters are more similar, are placed in one of them to determine the level of their damage based on the cluster in which they are located. In the motion method, a tree is created using customer data based on "if-then" rules, and then new customers enter from the root node to reach the leaf node. In this part, according to the characteristics of that node, the level of customer damage can be predicted and customers can be simulated from 730 to 7300 customers.

Statistical population and statistical sample: The statistical population used in this study is data related to one million fire insurance customers from Iran Insurance Company, which has been collected during the last five years and amounts to five million records. These data are stored in the Central Insurance Database and includes the customer's personal information and customer damage information. The total number of customers with a history of damage is about 70,000 customers, which is placed in a separate table. Also, for some operations such as exploratory analysis, samples of data are generated by reduction sampling method. Finally, the data of this research is 730 daily damages from Iran Insurance Company.

Optimized VAR simulation method

The Monte Carlo simulation method is one of the powerful tools in risk analysis. In this method, it is not necessary to assume that the return distribution is normal. Unlike the historical simulation method, the Monte Carlo simulation method does not use historical information, but in this method.

Theoretical model and feedback test (LR)

According to the proposed definition, the calculation of VAR based on a normal distribution will be as follows:

$$VaR(\alpha.t) = \mu - Z_a \sigma\sqrt{t}$$

μ and σ: average loss over a specified period of time and standard deviation

a: amount of space

t: time

A: Normal distribution

Another method used in this paper is the Brownian motion method. In order to generate random numbers in this method, it is assumed that the fluctuations in the amount of losses follow the Brownian motion fluctuations.

Another method used in this paper is the Brownian motion method, in order to generate random numbers in this method; it is assumed that the fluctuations in the amount of losses follow the Brownian motion fluctuations.

(2) P_T = P * exp [(μ-σ ^ 2/2) + σ_ε]

P_T: Damages incurred during T

P.: The amount of initial participation at time zero

μ: average

σ: Standard deviation from past fire damage

ε: Random number with normal distribution with mean zero and variance one.

3) The average variance is calculated as follows.

〖M_n〗 ^^= 1 / m ∑_ (i = 1) ^ mu_ (n-1) 〖σ_n〗 ^ 2 = 1 / m ∑_ (i = 1) ^ m 〖u ^ 2〗 _ (ni)

Where u_i = ln (P_i/P_ (i-1)). So that P_i is the amount of damage on day i and P_ (i-1) is the amount of damage on day i-1. Preliminary estimates of mean and variance are obtained to calculate the VAR for every 730 days starting from day n. (From nth day to n + n 729th day) in such a way that the VAR of the 731rd day is given from the 730th day of the previous day (730-1) and the risky value of the 732nd day is given from the 730th day of the previous day. The value (730-2) is obtained for that. And the trend continues to reach 730 VAR.

In this study, 〖μ_1〗 ^^ averaged 730 numbers from 1 to 730 and 〖μ_2〗 ^^ averaged 730 numbers from 2 to 731 and similarly 〖μ_730〗 ^^ averaged 730 numbers from 731 to 1460. By placing 〖μ_n〗 ^^ and 〖σ_n〗 ^ 2 in Formula (2), 7300 random numbers are simulated using fire damage data.

Preliminary estimates of value at risk are obtained for every 730 days starting on day n (from day n to day 729 + n). The estimates are then compared with the actual amount of damage according to this formula.

(4) π ^^= n_1 / (n_1 + n_0)

n_1: The number of times the estimated VaR is less than the expected return.

n_0: The number of times VaR is estimated to be greater than efficiency.

The first value at risk is compared to the actual amount of damage on day 731 and the second value at risk is compared to the actual amount of damage on day 502, and the comparison continues to the 730th value at risk with the 1460th real amount of damage. . If the value of μ^ is equal to the amount of acceptable error α, the result is desirable, or in other words, if exactly 7 of the 730 values of damage are greater than the value at risk, the Monte Carlo simulation method for generating random numbers to generate value in The risk of day n + 730 continues. Otherwise, the variance obtained in Formula (3) must be multiplied by the adjustment rate calculated in Table (1) and the random sample generation and comparison steps as

repeated above must be repeated until Where the equation μ = 0.01 is established. If μ /0 0.01, the adjustment coefficient continues to increase by increasing, and if μ> 0.01, by decreasing the variance resulting from Equation (3) until the production of the optimal VAR continues until μ 01 0.01. The increase or decrease of variance is represented by μ relation. This can be justified according to Table (1) and the rates obtained in the last column as the coefficient of variance adjustment.

In this table, the rates related to the variance adjustment coefficient are shown. The adjustment coefficient is determined by the rates to create the optimal results in the feedback test. π ﻪﺑ is calculated according to formula (4).

Table 1: the rates and quantities of damage recorded in the function

Rates	Function π̈ in %	Quantity of Damage
1	0.3	0.907
2	0.5	0.977
3	0.7	1.026
4	0.9	1.066
5	1.1	1.1
6	1.3	1.131
7	1.5	1.159
8	1.7	1.185
9	1.9	1.209
10	2.1	1.233

Feedback Test (LR)

Feedback test is used to evaluate the accuracy of the model used in calculating the value at risk. The null hypothesis presented in this fog test by Kopik states that the probability of failure or event of exceptions in action (π) is equal to the probability level considered in model (α). The test statistics show the likelihood ratio with LR and are calculated as follows:

(5) LR = -2in (α ^ (n_ (1 (1-α) ^ (n_0))) / π^^ (⟦n1 (1-π^) ^ (n_0) _))

This test has a chi-square distribution with one degree of freedom (Christoffersen, 2002). Any method used to calculate VaR can be controlled using the feedback test method. This test includes measuring the performance of VaR estimates in the past (Saeedi and Rai 2004).

Results and Discussion

Data Mining: Internal evaluation method is used to evaluate the results obtained in the data mining section. In internal evaluation, the model created by data mining methods is tested to measure its accuracy.

Internal evaluation is set up to verify that the parameters are appropriate for the intended purpose. Two methods are used for internal evaluation:

1- Evaluation with training and testing data

In the first method, to check how reliable the created model is and can predict the target variable, the data is divided into two categories: training and testing. The method is that first, the model is created using training data and then the model made with test data is tested to obtain its performance accuracy. This method is very suitable for testing predictive models (Han and Kamber, 2006)

2- Evaluation by light shadow method

The light shadow area method is used to evaluate the clusters. One of the drawbacks of the K-means fan is that it cannot offer the optimal K value. The light shadow method can be used to determine the optimal K value or to evaluate whether the clustering is well done.

Part II: Value at Risk

The data of this research includes 5000 amounts of daily damages from different branches of Iran Insurance and it is assumed that the source of compensation is the company's branches and the fluctuations in the amount of damages within each branch are independent of each other and follow

Brownian motion. In this section, first the nature of the damage distribution is examined and then the value at risk is calculated by estimating the adjusted amount. Finally, the VaR validity is measured. First, the normality of the distribution of 5000 daily damages and then the changes and fluctuations of the damages were determined, indicating that there is a sudden change from 2256 to 3255, and on other days there is a significant change. After determining the nature of the distribution of damage values, 1000 primary data of these values were used to estimate VaR and achieve the appropriate adjustment coefficient. Calculation of 500 means and variance using 1000 data was performed by the mentioned method and then 500 VAR were estimated. After calculating 500 values at risk, the in-sample test was performed with respect to π, and then adjusted 5 times and at the same time, 3000 times of simulation repetition and in each repetition, 2500000 random samples were produced and, therefore, 5 of the VAR was adjusted and finally, in the sixth time, the best VAR was obtained. After in-sample test, feedback test was used to evaluate the obtained estimates. After calculating the adjustment coefficient, an out-of-sample feedback test was performed for 4000 samples and finally the simulation results were obtained with its adjusted method by considering the coefficient estimated at 1.86 in Table 1. In order to show the efficiency of the adjusted method in cases where the amount of claims decreases or increases suddenly, a series of claims that continuously show the amount of decrease or increase in the number of claims was used as a subset.

Table 1: results of feedback

Test results are shown for a total of 4000 daily data and for a subset containing 1000 days. The time horizon is 1 day, 500 days of data is used with a daily damage review period of 5000. The number of times the amount of damage exceeds the value at risk is shown. As shown in Table 1, for 1% of the adjusted VAR for 4000 samples, the value of 67 is obtained as the number of times the damage exceeded the relevant VAR. In the case of 1000 data with the number of losses from the VAR is 12; the feedback test is 1.46 in significance level 0.33 for the conditions. (Table 2)

Table 2: the results of feedback test

Optimized VaR	Sample volume	Quantities of damage	% of damage	Verified percentage	Significance level
All samples	4000	67	0.016	29.61	0.00
Sample days 2256-3255	1000	12	0.012	1.46	0.33

Conclusion

The results of data mining operations are based on fire insurance data. Based on the results of this stage, it was observed that in addition to the appearance characteristics of fire insurance position such as volume, number of area, capacity, year of construction, and type of use, customer behavioral characteristics or their demographic characteristics also have a major impact on predicting the level of insurance customers. They have. This indicates that the parameters used for insurance pricing need to be reviewed and paid more attention to customer history, behavior, and damages. In fact, there is a huge shortcoming in determining the insurance tariff, which is currently based only on the characteristics of the insured subject. Of course, this does not mean that these features are not effective in determining the insurance tariff, but it does mean that these features are incomplete and cannot have the right result alone. Because these characteristics can describe the level of risk of customers and also collecting this information from customers is simple and not prohibited by law. Also, using the data of Iran Insurance Company, the VAR or in other words, the amount of capital required to cover the daily losses of fire insurance has been calculated using the average of 500 and the variance of the damage.

Optimized VAR was used to prevent market risk fluctuations and reduce it as much as possible. The optimized VAR was obtained during 6 repetitions and after 5 stages of variance adjustment, the results show that the adjusted VAR fog estimates the capital requirement by using the feedback test technique with a high probability.

References

[1] Efthymia Nikita, Panos Nikitas, (2020), Chapter 2.3: Data mining and decision trees, Statistics and Probability in Forensic Anthropology, 87-105
[2] Gregory Ricem, Tony Wirjanto, Yuqian Zhao, (2020), Forecasting value at risk with intra-day return curves, International Journal of Forecasting 36, 1023-1038
[3] Hameed Nazia, Antesar M. Shabut, Miltu K. Ghosh, M. A. Hossain, (2020), Multi-class multi-level classification algorithm for skin lesions classification using machine learning techniques, Expert Systems with Applications 141, 112961
[4] Hosseinzadeh Leila, Shaban Elahi, (2007), Classification of target customers in the insurance industry using data mining. Master Thesis in Information Technology Management, Tarbiat Modares University [In Persian]
[5] Jabbar Abdul, Pervaiz Akhtar, Samir Dani, (2019), Real-time big data processing for instantaneous marketing decisions: A problematization approach, Industrial Marketing Management 7, In press
[6] Libai Barak, Yakov Bart, Sonja Gensler, Charles F.Hofacker, Andreas Kaplan, Kim Kötterheinrich, Eike Benjamin Kroll, (2020), Brave New World? On AI and the Management of Customer Relationships, Journal of Interactive Marketing 51, 44-56
[7] Liyanage D. N. S. S. , Fernando G. V. M. P. A. , Arachchi D. D. M. M. , Karunathilaka R. D. D. T. , Perera, A. S. (2017), Utilizing Intel Advanced Vector Extensions for Monte Carlo Simulation based Value at Risk Computation, Procedia Computer Science 108, 626-634
[8] Peykarjoo, Haryar, Khosravi, (2006), investigating the risk of issuance in insurance companies using the value-at-risk method Insurance Industry Quarterly, Vol. 21 [In Persian]
[9] Saeedi and Rai, (2004), Fundamentals of Financial Engineering and Risk Management, Tehran: Samat Publications [In Persian]
[10] Shafiq Muhammad, Zhihong Tian, Ali Kashif Bashir, Alireza Jolfaei, Xiangzhan Yu, (2020), Data mining and machine learning methods for sustainable smart cities traffic classification: A survey, Sustainable Cities and Society 60, 102177
[11] Shahriar, Ahmadi, (2008), Calculating the amount and share of reinsurance maintenance in insurance companies with a risk-value approach [In Persian]
[12] Silahli Baykar, Kemal Dincer Dingec, Atilla Cifter, Nezir Aydin, (2019) Portfolio value-at-risk with two-sided Weibull distribution: Evidence from cryptocurrency markets, Finance Research Letters, in press
[13] Trung H. Le, (2020), Forecasting value at risk and expected shortfall with mixed data sampling, International Journal of Forecasting 36 , 1362-1379
[14] Vafaee Yeganeh Mohammad, Bahman Yasbolaghi SHarahi, Esfandyar Mohammadi, Fatemeh Havas Beigi, (2014), A Survey of the Relationship between Intellectual Capital and Performance of the Private Insurance Companies of Iran, Procedia - Social and Behavioral Sciences 114, 699-705
[15] Zaroni Hebert, Letícia B. Maciel, Diego B. Carvalho, Edson de O. Pamplona, (2019), Monte Carlo Simulation approach for economic risk analysis of an emergency energy generation system, Energy 172, 498-508.