International Journal of Engineering & Technology, 7 (4.5) (2018) 159-162

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Review On Application of Data Mining in Life Insurance

Vaibhav A. Hiwase[1]*, Dr Avinash J Agrawal[2]

[1,2] Shri Ramdeobaba College of Engineering and Management, Nagpur, Maharashtra, India
**Corresponding author E-mail:[1]hiwaseva@rknec.edu. [2]agrawalaj@rknec.edu*

Abstract

The growth of life insurance has been mainly depending on the risk of insured people. These risks are unevenly distributed among the people which can be captured from different characteristics and lifestyle. These unknown distribution needs to be analyzed from historical data and use for underwriting and policy-making in life insurance industry. Traditionally risk is calculated from selected features known as risk factors but today it becomes important to know these risk factors in high dimensional feature space. Clustering in high dimensional feature is a challenging task mainly because of the curse of dimensionality and noisy features. Hence the use of data mining and machine learning techniques should experiment to see some interesting pattern and behaviour. This will help life insurance company to protect from financial loss to the insured person and company as well. This paper focuses on analyzing hidden correlation among features and use it for risk calculation of an individual customer.

Keywords: *adverse selection; data mining; life insurance; risk factor; the null hypothesis*

1. Introduction

The goal of life insurance as a monitory provision and protection from financial loss to the insured person helps the community. Insurance operates through the risk of insured life. Hence profit and loss to the life insurance company mainly decided by the risk of insured people. Data mining and machine learning support various methods and technique to understand the behaviour of human nature from data. This behaviour helps to categories people into community possessing some common characteristics. These characteristics are directly and indirectly participating to represent some commonalities known as patterns. Thus pattern recognition becomes necessary not only to understand data but also create a model which automate the process of classification. Life insurance industry can use this model for underwriting and policy-making.

Adverse selection in life insurance industry is a situation where the purchase of insurance policy is affected by the asymmetric information. Asymmetry in information collection sometimes results in incorrect classification. As a result, low-risk customer along with high-risk customer can found to claim more often. This may cause a loss to the company. To recover from this loss insurance company may increase the premium rate. Hence customers can leave the policy if they were not able to pay the premium and may go for some different life insurance policy services with the low premium rate. Generally, low-risk customer has lower the premium because of lesser risk they are involved in. When these customers leave then the high-risk customer who really want the insurance benefit, being aware of their high risk they are involved in, are left behind in the insurance company which is a bad signal because they can claim more often. If this cycle continues company fails to get profit. Hence company needs a significant counterforce to protect against this kind of adverse selection. This counterforce can be created by a legislative measure which creates a compulsion measure on the market or by calculation of risk on an individual basis. Life insurance company gets a lot of information from their customers. Modern data mining can provide a useful technology to understand hidden knowledge which is useful to categories people purchasing insurance policy into the right risk group. This will ensure that company will not put the high-risk customer into a low-risk category and in effect, the company can take necessary action to enhance the loyalty of purchaser.

According to (Ansari, 2016), customer relationship management (CRM) helps to establish a long-term relationship with customers by delivering a great service and values to policyholders. This can be done by customer segmentation or classification which divide them into smaller groups, such that customers in the same group have similar characteristics. Ideally, organizations should have a good understanding about all of their customers, but this is not feasible in the real. Customer classification and clustering enable the firms to group similar customers together and help managers to better understand the customers' needs; because it is much easier to identify and analyze the characteristics of groups of customers rather than studying each customer individually as suggested by the author. After identifying these types of customers, the firm should motivate them to establish long-term relations. Also the customer loyalty will be enhanced, customer life-cycle will be optimized, and eventually, the firm will become more profitable. By dividing customers into different clusters firm can easily decide their insurance rate and also develop the effective marketing strategy. When the quality of clusters is not good then an expert advice is taken who uses some kind of interpretation. In high dimensional features space, this becomes a challenging task as various noise can be present in data. Also, curse of dimensionality may lead to rejection of various model. Hence it becomes important to test data with the various model which can reduce or at the best eliminate the curse of dimensionality. This review paper proposed a model to see and test the effects of high dimensional feature space in life insurance sector.

2. Related Work

Age and disability are key risk factors that cannot be readily substituted by alternative risk factors for many types of insurance in achieving the goal of providing affordable and accessible insurance. Risk-rated individual life insurer currently segmented in risk group by Age, Disability, Occupation, Leisure pursuits, Amount and duration of cover, Education, income, dwelling location, Behavioral habits (like smoking, drinking, drugs) ("Use of Age and Disability as Rating Factors in Insurance: Why Are They Used and What Would Be the Implications of Restricting Their Use?, position paper.", 2011). This factor though considered, as important factors may still not give any guarantee for classifying customers into the right risk group. So (Kahane, 2007) used the excess of 200 attributes which further classified into four groups to build a predictive model which is useful for underwriting and rate making. They use a two-stage approach, which involves survival analysis, and linear regression model, which helps them to estimate the risk level of each customer and the proneness to file a claim. Their results show that riskiest people on an average 12 times more expensive than least risky people in motor vehicles insurance. This is a clear indication that excess of 200 attributes helps them for correct classifications of the people. One of the biggest problems in working with high dimensional data is the presence of irrelevant features in a cluster and correlation of relevant features in different clustering. This is the main challenge in clustering can be thought as a curse of dimensionality. (Zimek, 2008) suggest some of the interesting approaches for solving curse of dimensionality, which can be tackled individually, or in the combination with a different approach. On the other hand, (Devi, 2016) examine how to get meaningful input variables for creating a model to do that. They use rule-based cluster model for motor policies.

(Ostaszewski, 1995) have explained in their research work a method of pattern recognition for risk and claim classification. They also made application to classify claims with regard to their suspected fraud content. Their result shows that fuzzy clustering is a valuable addition to the method of risk and claim classification. (Lemaire, 1990) aim to present the basic concepts of the fuzzy set theory in an insurance framework. The basic definition of fuzzy logic is presented and applied to provide a flexible definition of a "preferred policyholder" in life insurance. (Bezdek, 1984) has discussed in his fuzzy c-means (FCM) clustering program that the data to be analyzed must be in the form of numerical vectors called feature vectors, and the number of clusters must be predefined for obtaining the membership values of the feature vectors. Although fuzzy c-means clustering algorithm is relatively easy to implement and also have a low runtime it is not efficient due to the sensitivity of initial centroid values and the possibility of being trapped into a local optimum. (Stetco, 2015) improves the effectiveness and speed of Fuzzy C-means by utilizing the seeding mechanism of the K-means++ algorithm.

The main question is how to identify the key clients of the life insurance organization and how to analyze their behavioural attributes? (Madeira, 2002) compares Logistic regression, neural networks, decision trees and fuzzy modeling techniques by using cross validation measures for Target Selection. The four techniques are applied based on recency, frequency and monetary (RFM) value measures. Their result shows that Fuzzy modeling is slightly better with less standard deviation, using a much smaller number of variables. (Ansari, 2016) combine the fuzzy c-means clustering and genetic algorithms to cluster the customers of the steel industry. Their objective is to identify key customers and retain them. Their customers were divided into two clusters by using the variables of the LRFM (length, recency, frequency, monetary value) model. The customers in the first group considered as loyal customers and customers in the second group are considered as newcomers. Thus key client selection can be taken from the first group.

Although fuzzy logic in clustering with or without genetic algorithm seems to perform well in the insurance industry, it is still not gives an optimal result in asymmetric information. Information is asymmetric when insurance buyers have private information about their risk type and are not correlated with insurance demand. When this risk experience is not exposed to the insurance company it is not used to price insurance policy. (Finkelstein, 2014) tried to use asymmetric information and try to identify individual characteristics that are risk relevant and correlated with insurance demand, but still unused by insurance companies. Their results show that political economy of insurance regulation may play an important role in determining pricing function and so we can expect to find some other interesting knowledge. On the other hand, (Rahman, 2017) apply attribute selection techniques to properly classify the data and prove that classification techniques are very useful in classifying customers according to their attributes. Also (Qu, 2017) presented that the association rule based problem transformation method for multi-label feature selection in the framework of fuzzy-rough sets. In this method, the typical problems in multi-label classification is addressed, such as reducing the combination label number so can be used for selecting relevant attributes using their proposed model. (Kang, 2018) present new feature selection algorithms for aggregate data analysis. Although they focus on linear regression models for a continuous response, an extension to non-continuous response variable by logistic regression is possible in their algorithm.

3. Proposed System

Proposed system work is focused on generating the null hypothesis (H_0) to see how different features and their combinations are correlated with the risk factor. This features may not possess some relevancy and correlation among themselves due to noisy behaviour hence irrelevant features cannot discard in advanced. Also, the system can try to figure out a possible way to deal with asymmetric information when policyholder's risk is not exposed to the insurance company. Different data mining techniques and methods appropriate for features can be useful for accurate classification and clustering to understand some hidden interesting behaviour within data. This behaviour may be useful to derive new null hypothesis and research then can divert to prove it wrong. This process will continue until best or acceptable results will obtain, and also the hypothesis for meaningful features will be found. This will help to understand risk factor in life insurance.

The proposed method can validate null hypothesis by minimizing two main types of error known as Type I error and Type II error. Type I error rejects the true null hypothesis giving false positive results whereas type II error accept the null hypothesis when it is incorrect giving as a false negative in confusion matrix. The probability of making type I error or significance level is denoted by the Greek letter α (alpha) and is also called the alpha level. The alpha value is decided by researcher and hence the responsibility of making type I error solely lies on the shoulder of researchers. Often significance level is set to 0.05 (5%), but it mainly depends on the accuracy we want to obtain on the null hypothesis. High accuracy in acceptance of null hypothesis is obtained from the lower value of alpha level which is set to 0.01 (1%). p-value or probability value gives the smallest level of significance at which we can still reject the null hypothesis. Closer to 0 p-value represents more significant results. The probability of making type II error is denoted by the Greek letter β (beta). β depends mainly on sample size (n) and population variance (σ). Power of a test (which equals 1−β) is the probability of rejecting the false null hypothesis is the goal of this research. Generally, the power of the

test increases by increasing the sample size. In a two-sided test, the significant level alpha is equally divided between two sides of the normal distribution. This test helps to understand the rejection of null hypothesis on the different side of normal distribution and thus useful to know the side of hypothesis where it belongs. On a one-tailed test, rejection region will only be on the one side with a significant level of alpha (α). This test is thus used to test all-region on the one side of the hypothesis. We only test null hypothesis and the alternative hypothesis (H_1 or H_A) will be everything except a null hypothesis.

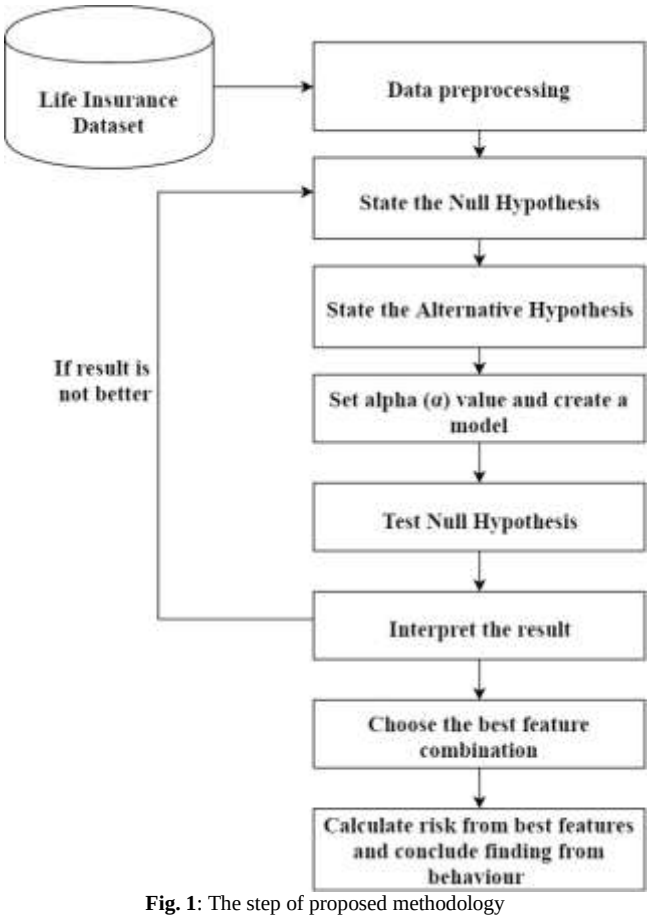

Fig. 1: The step of proposed methodology

Step1: Data preprocessing:
Real-world data generally inconsistent, incomplete and lacking in certain behaviour and trends. Hence preprocessing is necessary. Data goes through series of steps during preprocessing like data cleaning, data integration, data transformation, data reduction and data discretization.

Step2: State the Null Hypothesis:
Create and select meaningful features to find their correlation with response variable of risk. Use different feature combination and assumption if required and generate a null hypothesis. Relax on assumption.

Step3: State the Alternative Hypothesis:
The reason for to state the alternative hypothesis is that if the null hypothesis is rejected then there are many possibilities which need to be evaluated.

Step4: Set alpha (α) value and create a model:
Alpha value should be decided depending upon the accuracy and selection of model. Practice with log transformation may be useful on data for creation of appropriate data mining or machine learning model to satisfy proposed hypothesis.

Step5: Test Null Hypothesis:
Use model to run the tests of null hypothesis keeping the power of test close to 1. This will ensure we will not accept the false negative hypothesis. If the null hypothesis is accepted, then go for next meaningful features of step 1 otherwise evaluate rejection of null hypothesis on the different side of normal distribution and recreate the model or feature combination.

Step6: Interpret the result:
After testing null hypothesis, its performance is evaluated by calculating correlation of selected feature or feature combination with response variable of risk. If the result does not show strong correlation, then go to step2 to modify the null hypothesis by using knowledge of known behaviour of already tested results.

Step7: Choose the best feature combinations:
Continue the above loop till best group of features combination is found and generate an appropriate model which accurately classify policyholders into the correct risk category.

Step8: Conclude finding:
Calculate risk of each and every individual person and understand common behaviour and pattern in the dataset to conclude finding.

4. Conclusion

Traditionally segmentation was used in life insurance with the assumption that all individuals in the segment are "alike". The problem was small segments suffers from lack of statistical significance and poor prediction whereas large segments may not be homogenous for decision-making. Clustering had a humble origin in life insurance and its application seems to give promising results. Today, it is important to know the behavior of clustering in high dimensional feature space as insurance industry needs to be more accurate to perform well in the competition. In addition, adverse selection is a critical situation caused due to asymmetric information in the database and when insurance buyer has private information about their risk type and is not shared with Life Insurance Company. To understand the effect and behavior of features in high dimensional feature space, we cannot rely only on clustering. Hence, various data mining and machine learning techniques need to experiment to find out the correlation between the different groups of feature vectors with risk. This will enable the firm to understand the risk factors of higher dimensional feature space so that can be used to accurately predict the risk of policyholder on an individual basis in life insurance company as the counterforce against adverse selection and use it for enhancing the loyalty of policyholder and profit as well.

References

[1] (2011). *"Use of Age and Disability as Rating Factors in Insurance: Why Are They Used and What Would Be the Implications of Restricting Their Use?, position paper."*. Européen, Groupe Consultatif Actuariel.
[2] Ansari, A. a. (2016). "Customer clustering using a combination of fuzzy c-means and genetic algorithms." *International Journal of Business and Management* 11.7, 59.
[3] Bezdek, J. C. (1984). "FCM: The fuzzy c-means clustering algorithm.". *Computers & Geosciences* 10.2-3, 191-203.
[4] Devi, O. (2016). "Portfolio rule- based clustering at automobile insurance in Portugal. Diss". APA.
[5] Finkelstein, A. a. (2014). "Testing for asymmetric information using "unused observables" in insurance markets: Evidence from the UK annuity market.". *Journal of Risk and Insurance* 81.4, 709-734.
[6] Kahane, Y. e. (2007). "Applying data mining technology for insurance rate making: an example of automobile insurance.". *Asia-Pacific Journal of Risk and Insurance 2.1.*

[7] Kang, S. J. (2018). "Feature selection for continuous aggregate response and its application to auto insurance data." *Expert Systems with Applications* 93, 104-117.
[8] Lemaire, J. (1990). "Fuzzy insurance.". ASTIN Bulletin: *The Journal of the IAA* 20.1, 33-55.
[9] Madeira, S. a. (2002). "Comparison of target selection methods in direct marketing." *European Symposium on Intelligent Technologies, Hybrid Systems and their implementation on Smart Adaptive Systems.*
[10] Ostaszewski, R. A. (1995). "Fuzzy Techniques of Pattern Recognition in Risk and Claim Classification". *Journal of Risk and Insurance*, vol. 62, Issue 3, 447-482.
[11] Qu, Y. e. (2017). "Associated multi-label fuzzy-rough feature selection.". *Fuzzy Systems Association and 9th International Conference on Soft Computing and Intelligent Systems (IFSA-SCIS)*, 2017 Joint 17th World Congress of International. IEEE.
[12] Rahman, M. S. (2017). "Analyzing Life Insurance Data with Different Classification Techniques for Customers' Behavior Analysis.". *Advanced Topics in Intelligent Information and Database Systems*. Springer International Publishing, 15-25.
[13] Stetco, A. X.-J. (2015). "Fuzzy C-means++: fuzzy C-means with effective seeding initialization." *Expert Systems with Applications* 42.21, 7541-7548.
[14] Zimek, A. (2008). "Correlation Clustering". Dissertation, LMU München: Faculty of Mathematics, Computer Science, and Statistics.
[15] Dixit, S., & Agrawal, A. J. (2013). Survey on review spam detection. *Int J Comput Commun Technol* ISSN (PRINT), 4, 0975-7449.
[16] Roiger, R. J. (2017). Data mining: a tutorial-based primer. CRC Press.