Support Vector Machines for Classifying Policyholders Satisfactorily in Automobile Insurance

Zuherman Rustam\(^1\) and Ni Putu Ayu Audia Ariantari\(^2\)

\(^1\)Department of Mathematics, Universitas Indonesia, Depok, 16424, Indonesia, \\
\(^2\)Department of Mathematics, Universitas Indonesia, Depok, 16424, Indonesia, \\
rustam@ui.ac.id

Abstract: In every insurance company, the satisfactorily of policyholder is important to predict the future of the company. This leads to the point that one needs a system to classify policyholders satisfactorily. In this study, we proposed the used of machine learning, which is Support Vector Machines, to classify policyholders satisfactorily. Thus, we also need to focus on car policyholders’ policy. These will be the risk factors that one will classify using Support Vector Machines. By defining risk, insurance company can predict uncertain events easier. If the selected risk factors are adequate then it will ensure the sustainability of the insurance company by avoiding bankruptcy. Hence, several risk factors need to be employed to gain a good explanatory for classifying the policies. It will result in empirical evidence that every insurance company desired most to improve their bottom line. Therefore, Support Vector Machines is claimed to result in a reliable data to classify policyholders satisfactorily.

1. Introduction

Insurance is agreement between two parties, which are insurance company and policyholder, that become a prior for insurance company to receive premium from policyholder. In return, insurance company will provide replacement to policyholder for their loss, damage, incurred costs, lost profits, or legal liability to the third party that might be occurred because of uncertain events. Insurance company will also provide payment towards policyholder’ death based on the agreement. There are several advantages of insurance in a country, such as risk transfer, risk based pricing, and investment function.

Nowadays in Indonesia, insurance companies are developed align with the economic growth. It also means that insurance market competence is currently increasing. This is one of the reason why having an adequate risk factors are important for insurance company. The other reason is the accident rate has increase significantly over the past four years. Automobile insurance need to do an adjustment for the premium since the company need to be prepared for the probability of high indemnities. Hence, automobile insurance can assure their own economic sustainability by avoiding bankruptcy.

On this matter, one will focus on risk in Automobile Insurance. Automobile Insurance is a part of general insurance. On a daily routine business activities, insurance company will always prepare for a sudden incurred cost in near future. It is important for insurance company to analyst their data to gain a future prediction. This is one of the preventive action to encounter bankruptcy. Defining risk itself is important for insurance company to predict the uncertain events of future claims. Variables that one used will be correlated with claim rates to predict future claims. Hence, classifying risk factors is
needed to maintain operational efficiency. It is also a way to keep the customers by classifying their satisfactorily.

In this paper, one proposed the used of machine learning, which is Support Vector Machines, to classify the risk factors. Support Vector Machines is a well-known machine learning for its excellence to classify dataset. The proposed method in this paper has not been applied to this problem before even though there are several journals that has applied support vector machines in insurance company. This paper is elaboration from M. J. Segovia Vargas, M. Camacho Minano, D. Pascual Ezama. in Risk factor selection in automobile insurance policies: a way to improve the bottom line of insurance company. One will focus on real problems that Indonesian Insurance company has been through. D. Richadeu, in Automobile Insurance contracts and risk of accident: An empirical test using French individual data and S. Salcedo Sanz, J. L. Fernandez Villacanas, M. J. Segovia Vargas, C. Bousono Calzon in Genetic programming for the prediction of insolvency in non-life insurance companies, has discusses about the test needed to gain empirical evidence that one also will do in this paper. Furthermore, S. Salcedo Sanz, M. Prado Cumplido, M. J. Segovia Vargas, F. Perez Cruz, C. Bousono Calzon in Feature Selection methods involving Support Vector Machines for prediction of insolvency in non-life insurance companies and M. J. Segovia Vargas, S. Salcedo Sanz, C. Bousono Calzon in Prediction of Insolvency in non-life insurance companies using support vector machines and genetic algorithms, has discussed about the use of support vector machines to predict insolvency while in this paper one will use multiclass support vector machines that is being modified with kernel trick, radial basis function, also using confusion matrix as classifier performance to improve the model proposed. Hence, it is considered as one’ main contribution in Insurance discipline. Therefore, machine learning is being employed to show empirical evidence for Automobile Insurance to improve their bottom line. In the next section, one will discussed about Support Vector Machines; Variable Selection, Data and Classifier Performance; Experiment Results; Conclusion and Future Research.[1-9].

2. Support Vector Machines

Support vector machines was invented by Vapnik in 1960s. The main field of study that being used to develop its algorithm is based on the statistical learning theory. In recent years, support vector machines is well known as one of the effective machine learning. It claimed as a method that has a high classification efficiency. It mostly used for learning classification and regression of the data. Support vector machines can be explained as an attempt to find the best hyperplane that can be used to classify two different class of data. The main objective of Support Vector Machines is to find a hyperplane which maximizes the distance of itself with the nearest data. Those distance is called margin. The bigger the margin is the more possibility to produce smaller error in generalization. This matter is being represented in Figure 1.
Support Vector Machines can be classified into two major type which are linear support vector machines and multiclass support vector machines. Based on real world problems, it is very rare to find a linear separable for the dataset. It is the reason why multiclass support vector machines is existed.

In Figure 2, given some data in two-dimension space. It shows that the data will appears random in such way that one needs more than a linear function to separate two class of data. In Figure 3 the data is divided into two parts by hyperplane. This is the main idea of multiclass Support Vector Machines. This matter can be changed into Quadratic Programming problems which aims to minimizing or maximizing objective function with some subject to the constrains as the consideration.

Let \( \{x_i, y_i\}_i^N \) is the dataset with N is the number of sample, \( x_i \in R^D \) is feature of vector starting from sample i-th, D is the number of dimension also known as feature, \( y_i \) is class label for \( x_i \). In this paper, one will use multiclass support vector machines. Thus, \( y_i \in \{1,2,3,\ldots,k\} \) with k as the number of class that one will used. The main formula of support vector machines is written as follow:

\[
f(x) = w \cdot x + b
\]

This formula is used to maximize the margin with \( w \) is a parameter for weight and \( b \) is a parameter for bias. As it shown in Figure 1., support vectors are the nearest data from the hyperplane. The margin is defined as:
\[
\frac{|f(x)|}{\|w\|} \quad (2)
\]

Equation (2) will be develop to generalize and produce smaller error as possible. This equation will be written as follow:

\[
\max_{w,b} \frac{|f(x)|}{\|w\|} \quad (3)
\]

This margin equation (3) will develop into Quadratic Programming Problem. This is because maximizing margin means minimizing \(\|w\|^2\). Hence, the new equation is shown as follow:

\[
\min \frac{1}{2} \|w\|^2 \quad (3)
\]

subject to \(y_i(w \cdot x_i + b) \geq 1\) \(i = 1, 2, ..., N\)

The equation (3) will be develop to be primal optimization of Support Vector Machines. It written as follow:

\[
\min \left( \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{N} \xi_i \right) \quad (4)
\]

subject to \(y_i(w \cdot x_i + b) \geq 1 - \xi_i\) \(i = 1, 2, ..., N\)
\(\xi_i \geq 0\) \(i = 1, 2, ..., N\)

In equation (4), \(\xi_i\) and \(C > 0\) are parameter to minimize error classification and maximizing the hyperplane. The smaller the value of \(C\) will maximize the margin. To find dual form, we will apply Lagrange Multiplier as follow \(L(w, b, \alpha) = \frac{1}{2} \|w\|^2 - \sum_{i=1}^{N} \alpha_i [y_i (w \cdot x_i + b) - 1]\) to equation (4). Then we will have dual form as follow

\[
\min \left( \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) - \sum_{i=1}^{N} \alpha_i \right) \quad (5)
\]

subject to \(\alpha_i \geq 0\)
\(\sum_{i=1}^{N} y_i \alpha_i = 0\)

In equation (2), Lagrange Multiplier is also can be written as follow

\[
L(\alpha_i) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j k(x_i, x_j). \quad (6)
\]

In equation (6), \(k(x_i, x_j)\) is a kernel function. One also aims to maximizes \(L(\alpha)\) from equation (3) by obtaining \(\alpha\). It will goes as follow \(w = \sum_{i=1}^{l} \alpha_i y_i x_i \) \(0 \leq \alpha_i \leq c\) and \(\sum_{i=1}^{l} \alpha_i y_i = 0\). It shows that one will use dual form of Support Vector Machines to utilize kernel trick for efficiency improvement. This study used radial basis function as kernel trick which can provide nonlinear classification. It will be written as follow

\[
k(x_i, x_j) = \exp\left(-\frac{||x_i - x_j||^2}{2\sigma^2}\right) \quad (7)
\]

3. Variable Selection, Data, and Classifier Performance

In this section, one will show variable selected, the dataset also the success determination of classifier performance. This study will use the dataset from one of the largest automobile Insurance in Indonesia. This dataset consists of 13,635 samples automobile insurance policies. Only some of the data can be taken due to industrial confidentially issues. As one already stated in Introduction, this paper will employ some variables as the risk factors. Variables that will be shown are the common
things that needed for insurance business policy. It is both qualitative and quantitative values for the company data.

The dataset will be divided into two parts which are training dataset and test dataset. The aims of training dataset are to build up a model for selection matters while test dataset is for validating the model that has been built from training dataset. It will be described in more specific way in the next section.

| Risk Factors  | Computer Codes | Explanation                                                                 |
|--------------|----------------|-----------------------------------------------------------------------------|
| Salutation   | SAL            | This variable will present information whether the policyholder is a male, female or even corporate |
| Region       | REG            | This variable will only consist of 47 regions in Indonesia. This regions that one use are based on the available data |
| Kind of Vehicle | KoV            | This variable is a categorize of vehicle type into four classes based on the number of its axles |
| Use          | USE            | This variable will consist of four values i.e. Private, Private Corporate, Corporate properties, Commercial used |
| Seats        | NoS            | Vehicle number of seats. This variable will consist of four values i.e. ≤2, 3-5, 6-10, > 10 |
| Claim        | CLA            | This variable is based on the historical claim that the policyholders do in one period. There are only two values which are Yes or No |

The samples are consisted of 2,056 samples that have a historical claim in policy within a certain period then the rest of it has no claim. This dataset use a year as a period. The details of the data will be seen as follow:

| Claim       | SAL | Frequency | Percentage |
|-------------|-----|-----------|------------|
| Had Claim   |     |           |            |
| Male        | 302 | 2.2%      |            |
| Female      | 93  | 0.7%      |            |
| Corporate   | 1661| 12.3%     |            |
| Total       | 2056| 15%       |            |
| Had No Claim|     |           |            |
| Male        | 1597| 11.7%     |            |
| Female      | 383 | 2.8%      |            |
| Corporate   | 9598| 70.3%     |            |
| Total       | 11578| 85%     |            |

Confusion matrix occupy important role for classifier performance. At this matter, confusion matrix taking parts to leads into result from the dataset that has been classified using support vector machines. A confusion matrix will present the result from classification model which are the number of correct and incorrect predictions compared to the actual outcomes from the dataset. The following table II will show confusion matrix the form explanation while table III will define each terms of confusion matrix.
Table 3. Confusion matrix form explanation

| Observed | Had Claim | Had No Claim |
|----------|-----------|--------------|
| Predicted | Correct had claim (True Positives) | Type II Error (False Negatives) |
| Had Claim | | |
| Had No Claim | Type I Error (False Positives) | Correct had no claim (True Negatives) |

Table 4. Definitions of confusion matrix terms

| Definition | True Positives (TP) | True Negatives (TN) | False Positives (FP) | False Negatives (FN) |
|------------|---------------------|---------------------|----------------------|----------------------|
| Definition | The number of claimed data that is classified as claimed | The number of no claimed data that is classified as no claimed | The number of no claimed data that is classified as claimed | The number of claimed data that is classified as no claimed |

The number of data that is true predicted is the amount of true positives and true negatives while the number of data that is false predicted is the amount of false positives and false negatives. The formula to find the entries of confusion matrix will be written as follow:

\[
\%TP = \frac{\#TP}{\%testdata} \times 100\%
\]
\[
\%TN = \frac{\#TN}{\%testdata} \times 100\%
\]
\[
\%FP = \frac{\#FP}{\%testdata} \times 100\%
\]
\[
\%FN = \frac{\#FN}{\%testdata} \times 100\%
\]

The entries of confusion matrix are in percentage. \(\%TP\) stands for the percentage for true prediction while \(\#TP\) means the number of true prediction. It means the number of the claimed data that is classified as claimed and so on.

Confusion matrix also takes parts to gain the accuracy rate. The higher the rate is the better classification performance. The proportion of the correct total number predictions is the accuracy which can be written as follow:

\[
accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\%
\]

4. Experiment Results
This study proposes the used of Support Vector Machines to classify the dataset given. It will use radial basis function for kernel trick in Support Vector Machines. The best result is the one with the best accuracy. It will be shown later in the table V and table VI.
Table 5. Results

| Training Data | % Accuracy | Running Time |
|---------------|------------|--------------|
| 10            | 32.17      | 3.64         |
| 20            | 74.13      | 1.83         |
| 30            | 76.32      | 41.27        |
| 40            | **84.08**  | **64.02**    |
| 50            | 78.35      | 73.61        |
| 60            | 73.50      | 96.97        |
| 70            | 68.50      | 110.50       |
| 80            | 55.38      | 127.30       |
| 90            | 78.75      | 136.70       |

In table V, the best accuracy is 84.08% with 40% of the data training and 64.02 for the running time. The type of kernel that one used is radial basis function with parameter 0.05. This result is considered as quite good because it is higher than 80%. With confusion matrix, one will be able to present the data clearly. The confusion matrix form will be presented as follow:

Table 6. Confusion matrix

| Had Claim | Had No Claim |
|-----------|--------------|
| Had Claim | 48.38        | 1.63         |
| Had No Claim | 14.29    | 35.71        |

Table VI showed the form of confusion matrix which provide the major information that is considered as empirical evidence for future claims. There is 48.38% of claimed data classified as claimed, 35.71% of no claimed data are classified as no claimed, 14.29% of no claimed data are classified as claimed, and 1.63% of claimed data are classified as no claimed. This means from the 40% of the training data taken for building up the model, the other 60% is a test dataset which were taken for validating the model that has been made using the training data. It means 60% from 13,635, which are 8,181, are considered as test dataset.

5. Conclusions
This study aims to gain a reliable empirical evidence that will be beneficial for future claim prediction. It will give the automobile insurance company an advantage to improve their bottom line. The proposed method, Support Vector Machines, had been occupied to solve this matter. Even kernel trick, radial basis function, also being used for a better efficiency of Support Vector Machines. After the one got the outcomes, it also shown the classifier performance by utilizing confusion matrix. The result of the accuracy is 84.08% which is quite satisfying. Only 1.63% of the claimed data are classified as no claimed. It means the classifier can be considered as efficient. The risk factors that one used in this research also adequate to ensure the sustainability of the automobile insurance company. Hence, the proposed method is recommended for future research towards automobile insurance. It will be better if it also contains feature selections to select the risk factors. Thus, the variable taken for the risk factors are assumed as the most significant risk factors to determine the better result later.

References
[1] M. J. Segovia Vargas, M. Camacho Minano, D. Pascual Ezama. 2015 Risk factor selection in automobile insurance policies: a way to improve the bottom line of insurance company. Review of Business Management, 17 1228-1245.
[2] B. Kramer. 1997 N.E.W.S.: A model for the evaluation of non-life insurance companies. European Journal of Operational Research, 98 419-430.

[3] D. Richadeu. 1999 Automobile Insurance contracts and risk of accident: An empirical test using French individual data. The Geneva Papers on Risk and Insurance Theory, 24 97-114.

[4] S. Salcedo Sanz, J. L. Fernandez Villacanas, M. J. Segovia Vargas, C. Bousono Calzon. 2005 Genetic programming for the prediction of insolvency in non-life insurance companies. Computers and Operations Research, 32 749-765.

[5] S. Salcedo Sanz, M. Prado Cumplido, M. J. Segovia Vargas, F. Perez Cruz, C. Bousono Calzon. 2004 Feature Selection methods involving Support Vector Machines for prediction of insolvency in non-life insurance companies. Intelligent System in Accounting, Finance and Management, 12 261-281

[6] M. J. Segovia Vargas, S. Salcedo Sanz, C. Bousono Calzon. 2004 Prediction of Insolvency in non-life insurance companies using support vector machines and genetic algorithms. Fuzzy Economic Review, 9 79-94.

[7] A.I. Dimitras, S.H. Zanakis, C. Zopoundis. 1996 A survey of businesss failures with an emphasis on prediction methods and industrial applications. European Journal of Operational Research, 90 487-513.

[8] B. Scholkopf, A. Smola. 2002 Learning with kernels. Cambrige, MA: MIT Press

[9] H. Nurmi, J. Kacprzyk, M. Fedrizzi. Probabilistic, fuzzy and rough 1994 concepts in social choice. European Journal of Operational Research, 95 264-277.