Software Engineering
2018; 6(4): 121-127
http://www.sciencepublishinggroup.com/j/se
doi: 10.11648/j.se.20180604.13
ISSN: 2376-8029 (Print); ISSN: 2376-8037 (Online)

Risk Assessment Predictive Modelling in Ethiopian Insurance Industry Using Data Mining

Sisay Wuyu, Patrick Cerna

Department of Information Technology, Federal TVET Institute - University, Addis Ababa, Ethiopia

Email address:
sisaywuyu2017@yahoo.com (S. Wuyu), pcerna@acm.org (P. Cerna)

To cite this article:
Sisay Wuyu, Patrick Cerna. Risk Assessment Predictive Modelling in Ethiopian Insurance Industry Using Data Mining. *Software Engineering.* Vol. 6, No. 4, 2018, pp. 121-127. doi: 10.11648/j.se.20180604.13

Received: June 25, 2018; **Accepted**: December 5, 2018; **Published**: January 17, 2019

Abstract: Risk management has long been a topic worth pursuing, and indeed several industries are based on its successful applications, insurance companies and banks being the most notable. Data Mining (DM) - is one of the most effective alternatives to extract knowledge from the great volume of data, discovering hidden relationships, patterns and generating rules to predict and correlate data, that can help the institutions in faster decision-making or, even reach a bigger degree of confidence. This research was conducted in a form of case study in the Ethiopian Insurance Corporation (EIC) at its main branch located at Legehar- Addis Ababa. The general objective of the study is to examine the potential of data mining tools and techniques in developing models that could help in the effort of Risk level pattern analysis with the aim of supporting insurance risk assessment activities at EIC. In this research two data mining technique which are decision tree and neural network. The best decision tree model, which is selected as a working model among the numerous models generated during the training phase, was able to correctly classify 75% percent of the 3100 policies in the validation data set. 96% of low-risk policies were correctly classified. Significant number of misclassification was observed on high risk level. The output of these experiments indicated that the classification task of records using the Risk level, both decision tree and neural network have performed with significant error. Decision tree has shown an accuracy rate of 75 percent while neural networks classified 58% records correctly. The overall performance of decision tree was better in classifying values than neural network.

Keywords: Data Mining, Risk Assessment, Decision Tree, Neural Network, Ethiopia

1. Introduction

Risk management has long been a topic worth pursuing, and indeed several industries are based on its successful applications, insurance companies and banks being the most notable. What gives this discipline enhanced attention and renewed prominence is the belief that nowadays we can do a better job of it. This perception is based on phenomenal developments in the area of data processing and data analysis. The challenge is to turn 'data' into information, knowledge and deep understanding [1]. Traditionally, actuaries develop risk models by segmenting large population of policies into risk groups, each with its own distinct risk characteristics. Premiums are then determined for each policy in a risk group based on the risk characteristics of the group, such as mean claim rate and mean claim severity, as well as on the cost structure of the company, its marketing strategy, competitive factors, etc. [2]. However, the basic tent in the industry is that no rating system can be perfect and competition therefore compels insurance companies to continually refine both the delineations they make among the risk groups and the premium they charge. The analytical methods employed by actuaries are based as much on statistical analysis as they are on experience, expert knowledge, and human insight [3].

Thus, it is widely recognized that any risk model one develops is likely to overestimate the true levels of risk of some groups of policies and underestimates the risks of others. In the last decades, in which most of the operations and activities of the public and private institutions are computationally registered and accumulate in large databases, the data mining technique - Data Mining (DM) - is one of the most effective alternatives to extract knowledge from the great volume of data, discovering hidden relationships, patterns and generating rules to predict and correlate data, that can help the institutions in faster decision-

making or, even reach a bigger degree of confidence [4]. The methods are technologies that exist, regardless of the data mining context, when applied in the KDD, they produce good results in the health area, changing data into useful knowledge and favoring the health practices based on evidence [5]. There are several methods, but the aim is not to exhaust the subject but to identify the most used. The main technologies are: Neural Networks, Decision Tree, Genetic Algorithms (AGs), Fuzzy logic and Statistics.

This research was conducted in a form of case study in the Ethiopian Insurance Corporation (EIC) at its main branch located at Legehar- Addis Ababa. The availability of required amount of data and the willingness of the management to co-operate in providing information necessary for this study encouraged the researcher to conduct this research at EIC. Ethiopian Insurance Corporation, the only government insurance company, was established in January 1st 1976. There are about 30 classes of business under life and non life category that EIC acts upon as a risk financing mechanism. One of these is the motor class. A typical motor policy is an amalgam of several different elements of cover and presents special problems for the insurer since each element impacts on a separate basic account: own vehicle, third party property damage, or third party personal injury (see chapter three for more detail). Moreover, it is the experience of the insurance company that motor class is characterized by relatively high claim frequency and cost. For these reasons, the researcher opted to experiment on motor insurance data.

However, an increasing size of the customer profile coupled with the small number of employees, made it difficult for the insurers, to regularly audit renewal acceptance. It is, therefore, with this understanding that this experimental research has been undertaken to develop a predictive model using data mining technology for the purpose of insurance risk assessment activities which is very important to keep the business safe and put ahead of its competitors. One alternative solution to this problem is to classify policyholders based on their degree of risk exposure using the portfolio of policy and claims data so that the insurer can easily identify those who demand priority attention from those that need little or no attention at the time of renewal. In this respect, data mining techniques could be effective tools in designing such a classification problem. With data mining, one can build a model that predicts the risk level of a policy by discovering highly complex relationships among the underwriting and claims data. The general objective of the study is to examine the potential of data mining tools and techniques in developing models that could help in the effort of Risk level pattern analysis with the aim of supporting insurance risk assessment activities at EIC.

2. Related Works

Stetco (2015) used Data Mining improves the effectiveness and speed of Fuzzy C-means by utilizing the seeding mechanism of the K-means++ algorithm, while 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.

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.

Similarly, 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. On the other hand, 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. While in recent studies 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. Materials and Methods

3.1. Data Collection

Selecting the best data for targeting model development requires a thorough understanding of the market and the objective. Although the tools are important, the data serves as the frame or information base. The model is only as good and relevant as the underlying data. The data employed in this research was collected from the Ethiopian Insurance corporation policy and claim database which contain millions of records. For the purpose of this research, motor policy records both policy and claim records were taken and all the necessary preprocessing tasks were carried out. The quality of the data is the most important factor to influence the quality of the results from any analysis. The data should be reliable and represent the defined target population. Data is often collected to answer specific questions using the

following types of studies. EIC has a data base to store all life and Nonlife policies related records. Motor policy records are found in Non life category therefore the researcher fined this data base as an appropriate source of information for this study.

This original database is stored and maintained separately from the KDD process. Control and access to this original database is restricted due to the confidential nature of the material of interest. The employees of the company then perform the data selection, based on input requirement from the researcher and domain expert suggestion. But accesses of some requested attributes from the researcher were restricted. As mentioned in the methodology section of chapter one, for the purpose of data collection, EIC was chosen due to large experience it has in the area and having large database implying a huge amount of records. EIC is dealing with 45 type of cover classified in to two main category i,e 15 of them are Life and the remaining 30 are None life cover. Motor policy is one of non life policy having a large amount of records.

In addition, from among the three types of cover i.e comprehensive, damage, third party, comprehensive cover was chosen for this research. This is because comprehensive cover constitutes major proportion of the total motor policy existing in the company. Moreover, the company has been dealing with relatively large number of motor claims from comprehensive cover. The other point that had to be addressed during data collection was despite 30 years experience in the business, the data employed for this research was collected policies with in past four years. The initial source dataset was extracted from the motor insurance portfolio database of EIC. The well experienced staffs in the IT department of the company extract all the required data from respective tables in the database and provide a soft copy in excel format. Next, the real and symbolic data fields (attributes) within each of the policy and claims databases had to be combined into a single database that could be used for decision tree and neural network training.

The database selected for this research consist more than 93000 motor policy records with 80 attributes and still there were a number of records stored manually waiting to be entered to the automated system. From the available 93000 records a sample 15000 were take for this study. The sampling technique employed here is random sampling taking 10 records from every 60 records.

3.2. Data Pre-Processing

The database being used for the purpose of this research work has a number of limitations. The limitations include missing values in various fields and encoding problem. At the time of this work only a limited subset of all possible features (or attributes) were available in sufficient numbers to allow investigation. In the event when there is a comprehensive automated data collection process, perhaps an optimal model may be produced. The motor policy record in policy and claim database at EIC contains more than 45 attributes and to decide on the relevant attributes for this study, a discussion was made with the domain expert at underwriting and claim unit at the corporation. Thus, records consisting of these attributes were preprocessed and prepared as stated in the subsequent sections before the data were provided to the decision tree algorithm. Then the decision tree algorithm calculates the information gain to select valid attributes and makes classification based on the selected attributes. This process involves summarization, data encoding, handling missing values, deriving new fields, and finally preparing the data into a form that is acceptable to the neural network [6].

Finally the number of records in the final working dataset, after removal of records during the preprocessing stage, was 14098. Table 1 shows final list of variables that have been used in this study. The final task at this stage was preparing the data into a form that is acceptable to the neural network. Neural networks accept values in the range of 0 to 1 or –1 to 1. Fortunately, the WEKA software that was used in this study has the facility to automatically transform values into a form that can be understood by the neural network, i.e., in the range of 0 to 1. In this respect, the values in the numeric fields are scaled down to the range of 0 to 1 using the maximum and minimum values within the field.

Table 1. *List of Attributes after preprocessing.*

No	Attribute	Meaning	Value	Type
1	usage	Purpose of the vehicle	Private Own good Own service Public service General cartage	Nominal
2	Make	Brand of Car	Toyota Yamaha Nissan Mitsubishi Suzuki Isuzu Hyundai Mercedes DAF IVECO Scania	Nominal

No	Attribute	Meaning	Value	Type
			Renault Mazda Tracker Ford Higher Chevrolet Peugeot Aeolus Volvo BMW Abay Honda Land rover Lifan Awash Nissan Sky bus Bajaj Motor cycle Pick up	
3	Body type	body type of the car	Automobile Bus Tanker	Nominal
4	CC	Carrying capacity of the car	<2000 2000-4000 4000-6000 6000-8000	
5	Agevh	Age of the vehicle	1-10 10-20 20-30 30<	
6	seatC	Seat capacity of the car	1-13 13-26 26-39 39-52 52<	Numeric
7	PLH	Policy holder	Male Female LI	Nominal
8	Ageplh	Age of the policy holder	Adolescent Adult Old	Numeric
9	Risk level	Level of the risk	High Medium Law	Nominal

4. Experimental Results

In this research two data mining technique which are decision tree and neural network. Decision tree is a method commonly used in data mining uses a decision tree to go from observations about an item to conclusions about the item's target value [7]. Neural network are powerful techniques for representing complex relationships between inputs and outputs based on the neural structure of the brain [8].

4.1. Decision Tree Model Building

The decision tree software employed for the purpose of this research was the WEKA software package, which contains several classifiers, clusters and association algorithms. The data preprocessed in the preceding steps are well suited to the WEKA software to train a classifier model. Here, the researcher was more interested in generating rules to explain risk exposure of policies and to come to an understanding of the most important factors affecting the risk level of a policy than in simply classifying particular policies as risky or not, or predicting which future policies would be risky. As a consequence, the decision tree that made the best predictions was not the one most useful for us. Instead, the one that generates sound rules was given priority in model selection.

With this understanding, numerous decision trees were constructed by adjusting different parameters taking 78-22 percent split of training and testing data, respectively, from the data file. For each decision tree, the corresponding rule sets were extracted. Finally, on the basis of evaluation of the rule sets made by domain experts, a tree model whose rule set is found to be meaningful was selected as a working model for further stages of the study.

In relation to this, J4.8 was the algorithm used to construct the decision trees in the experiments. The extraction of the corresponding rules, in the experiments, from the decision trees were managed using J4.8 PART. The summary of constructed decision trees using different parameter is depicted in Figure 1 the extracted rules are annexed for reference.

```
| | | | | | | make = NISSAN
| | | | | | | | age of v = (10-20]: Low (4.0)
| | | | | | | | age of v != (10-20]: High (20.0/1.0)
| | | | | | | make != NISSAN: Low (50.0/5.0)
| | | | | | | age != Adolescent: Low (2098.0/859.0)
| | | | | | type != Station Wagon
| | | | | | | make = AEOLUS
| | | | | | | | age of v = (1-10]: High (29.0/6.0)
| | | | | | | | age of v != (1-10]: Low (4.0)
| | | | | | | make != AEOLUS
| | | | | | | | make = TRAKKER
| | | | | | | | | age of v = (1-10]
| | | | | | | | | | phl = Legal entity
| | | | | | | | | | | usage = General Cartage
| | | | | | | | | | | | cc = (6000-8000]: High (67.0/18.0)
| | | | | | | | | | | | cc != (6000-8000]: Low (4.0)
| | | | | | | | | | | usage != General Cartage: High (12.0/3.0)
| | | | | | | | | | phl != Legal entity: Low (4.0)
| | | | | | | | | age of v != (1-10]: Low (7.0)
| | | | | | | | make != TRAKKER
| | | | | | | | | age of v = (20-31]
| | | | | | | | | | make = IVECO: High (52.0/17.0)
| | | | | | | | | | make != IVECO
| | | | | | | | | | | make = VOLVO: High (20.0/7.0)
| | | | | | | | | | | make != VOLVO
| | | | | | | | | | | | usage = General Cartage
| | | | | | | | | | | | | make = MERCEDES
| | | | | | | | | | | | | | cc = (2000-4000]: Medium
(39.0/23.0)
```

Figure 1. *Model structure of General Decision Tree with Pruning and with all attributes.*

4.2. Classification Results

The classification performance of the selected decision, from 8 models with different parameters tree were presented in Table 2 measured using the validation dataset. The Score column in the table shows the percentage of records that has been classified correctly.

Table 2. *Comparisons of Performance and Accuracy of Models.*

Type	General	Binary
Pruning	Yes	Yes
Attribute used	Reduced	Reduced
No. of leaves	1017	224
Size of tree	1141	447
Time (Sec)	0.25	0.77
CCI	74.80%	74.90%
ROC Area	0.806	0.86
AVG FPR	0.383	0.382
Recall	0.749	0.749

The results of the model for the validation dataset are given in Table 3. The output of this experiment depicted that out of the total test data (3100 records) provided 75 percent records were classified correctly. The results of the experiment has also shown that about 96 percent of the records in the class of Low risk were correctly classified while about 25 and 42 percent of the records in the medium and low crime classes respectively were classified correctly. Although several attempts were made, like re sampling the data set and SMOTE to balance the number of instances in a class, it was not possible to improve the performance of the classification above 75%.

Table 3. *Confusion Matrix for Selected Decision Tree.*

Actual	Prediction			Total	Score
	Low	High	Medium		
Low	1985	64	5	2054	96%
High	364	153	77	594	25%
Medium	241	19	189	449	42%
Total	2590	236	271	3097	75%

In the case of SMOTE (Synthetic Minority Oversampling Technique), for improving the accuracy over the minority classes, and wants to utilize boosting to maintain accuracy over the entire data set. The major goal is to better model the minority class in the data set, by providing the learner not only with the minority class instances that were misclassified in previous boosting iterations, but also with a broader representation of those instances. In this case, performance of the lower instance increases at the expense of the larger instance class. About 25% records are wrongly classified and this may account to the error made by the risk assessment experts in classifying the data employed in this study. The overall performance of the model is encouraging. The number of high risk policies that has been classified correctly is, however, small. One possible reason for the majority of high class classified as low could be during the assessment classification the researcher used claim frequency to categories risks due to lack of information on the severity of claim expressed in terms of total claim cost. This could lead to frequent miner claims may goes to higher class while a single but total lose claim goes to a low class while it is actually high in terms of claim cost.

4.3. Neural Network Model Building

The second data mining technique employed in this study is neural networks. To build the neural network model, WEKA software package is used and it employs back propagation algorithm in developing a model. In order to train a model there is usually a need to prepare the dataset in a form, which is suitable to the particular data mining technique and software. Neural networks technique can only process datasets when the values of attributes (variables) are numeric. Only the target class could be nominal valued variable. Moreover, the values of each attribute must be between 0 and 1 or between -1 and 1. Therefore, to make the dataset suitable for the neural network, values of all attributes were changed into numbers between the acceptable ranges (in this case between 0 and 1). Then the network was trained by presenting a set of instances repeatedly through all the training set addressing each instance in turn and making necessary corrections. When the entire list of records has

been presented, WEKA starts over at the beginning of the list and the training process was repeated until the termination condition of the iteration process is reached.

Using the neural network program (sub package) of the WEKA software, experiments were conducted i.e. classification of records on target classes of Risk level. In order to make comparison on the performance of decision tree and neural networks the same dataset provided to the decision tree is used for the neural network. The attributes ranked by information gain using attribute selection package in WEKA were provided to the neural network and hence training and testing of the model is based on these attributes. In conducting this experiment the parameters of the neural networks program were set to their default values. Accordingly by adjusting different parameters a number of neural network models were developed and the one performing best selected as depicted in Table 4.

Table 4. *Neural Network Confusion Matrix.*

Actual	Prediction			Total	Score
	Low	High	Medium		
Low	1599	447	8	2054	77.8%
High	291	223	83	597	37.3%
Medium	153	101	195	449	43.4%
Total	1943	871	286	3100	58%

The classification of records using the attribute Risk level resulted in significant number of records, which are incorrectly classified. Thus, in order to improve the performance of the model an attempt was made to modify some of the parameters. With this objective the researcher modified the number of hidden units to 6 and the learning rate (the rate at which the network weights were adjusted) and momentum (that tends to keep the change in the same direction from one iteration to the next) to 0.4 and 0.5 which were 0.3 and 0.2 in the default value respectively.

However, this trial has diminished the performance of the model and the accuracy rate obtained from this experiment was 58 percent. Moreover, an attempt was also made to set the number of hidden units at different level and other parameters such as learning rate and momentum are also modified. Nevertheless, it was not possible to improve the performance of the model in spit of all these efforts.

5. Conclusion

In this paper, an attempt was made to assess the potential applicability of data mining technology in support of risk assessment activity in the insurance industry. This experimental research, which employed the commonly used methodological approach in data mining researches, made use of two predictive modeling techniques, decision tree and neural networks, to address the problem. In the feature selection phase, more emphasis was given to the soundness of the rule sets extracted from the resulting decision trees. Accordingly, the better decision tree selected as a working model generates meaningful rules that would assign new policy records to the classes. Nine (9) attributes were selected in consultation with domain experts and ranked with decision according to their information gain. The best decision tree model, which is selected as a working model among the numerous models generated during the training phase, was able to correctly classify 75% percent of the 3100 policies in the validation data set. 96% of low-risk policies were correctly classified. Significant number of misclassification was observed on high risk level. Only 25% was classified correctly. But the resulting rule can meaningfully explain conditions for high risk exposure.

The output of these experiments indicated that the classification task of records using the Risk level, both decision tree and neural network have performed with significant error. Decision tree has shown an accuracy rate of 75 percent while neural networks classified 58% records correctly. The overall performance of decision tree was better in classifying values than neural network. Therefore, it is plausible to conclude that the decision tree data mining technique is more appropriate to this particular case than the neural network.

6. Recommendation

This paper has been conducted not only with academic contribution in mind but the findings of the research will help also initiate financial sectors to work on the application of data mining technology to gain competitive advantage in their field. Moreover, the research work can contribute a lot towards a comprehensive study in this area in the future. Data mining techniques could contribute a lot in assessing risks by identifying which particularities of a vehicle leads to certain risk level. Thus, it could be more important to use the data mining technique as a tool for the decision making process. In other words, EIC should employing data mining technology to support its risk assessment process which intern enables the corporation to sign effective premium rating.

The way in which risk incidents are designated as high, medium or low risk levels needs a revision. Some kind of risk may be grouped erroneously in a risk group that does not belong to it and lacks a sound justification. This may be one reason why the model has reported relatively a low accuracy rate in this particular experiment. Therefore, one can develop a model with a revised assessment classification following this recommendation. Although in this study encouraging results were obtained, during data collection the researcher was allowed to access only with 9 variables which could limit the risk factor. This leaves the model opened to be reviewed to include more risk factors such as detail claim information like type of claim, which may include medical expenses, repair and replacement

References

[1] Kennet, Y. a. (2001). Operational Risk Management: Apractical approach for data analysis.

[2] Apet, e. a. (1998). Insurance Risk modeling Using Data mining Technology.

[3] Dockrill, M. et al. (2001). Underwriting Management. Study Course 815. London: CII publishing Division.

[4] Cardoso ONP, Machado RTM. Knowledge management using data mining: a case study at the Federal University of Lavras. Rev Public Adm. 2008; 42 (3): 495-528.

[5] Rodrigues RJ. Information systems: the key to evidence-based health practice. Bull World Health Organ. 2000; 78(11): 1344-51.

[6] Stetco, A. X.-J. (2015). "Fuzzy C-means++: fuzzy C-means with effective seeding initialization." *Expert Systems with Applications* 42.21, 7541-7548.

[7] 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.*

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] Kang, S. J. (2018). "Feature selection for continuous aggregates response and its application to auto insurance data." *Expert Systems with Applications* 93, 104-117.

[13] Berry, M. a. and Linoff, G. (2000). Mastering Data mining: the art and science of Customer Relationship Management. New York: John Wiley & Sons, inc.

[14] Rokach, Lior; Maimon, O. (2008). Data mining with decision trees: theory and applications. World Scientific Pub Co Inc.

[15] Russell, S. & Norvig, P.: Artificial Intelligence: A Modern Approach. Prentice Hall, London (2003).