International Journal of Business and Economics Research
2014; 3(3): 128-134
Published online May 30, 2014 (http://www.sciencepublishinggroup.com/j/ijber)
doi: 10.11648/j.ijber.20140303.12

Car insurance risk assessment with data mining for an Iranian leading insurance company

Seyed Behnam Khakbaz*, Nastaran Hajiheydari, Marziyeh Pourestarabadi

Faculty of Management, University of Tehran, Tehran, Iran

Email address:

Khakbazbehnam@yahoo.com (S. B. Khakbaz), nhheidari@ut.ac.ir (N. Hajiheydari), m.pourestarabadi@gmail.com (M. Pourestarabadi)

To cite this article:

Seyed Behnam Khakbaz, Nastaran Hajiheydari, Marziyeh Pourestarabadi. Car Insurance Risk Assessment with Data Mining for an Iranian Leading Insurance Company. *International Journal of Business and Economics Research.* Vol. 3, No. 3, 2014, pp. 128-134.
doi: 10.11648/j.ijber.20140303.12

Abstract: Today's competitive market leads industry to a serious fight. This fight has guided some companies to a sightless selling. Insurance companies lose lots of money each year because of not profitable and risky customers which are attracted blindly. Risky customers are one of the most important treats to insurance companies; therefore some of these companies adopt a credit scoring and risk assessment approach for identifying profitable and risky customers. One of the most preferable methods for risk assessment is data mining. In this article, authors would demonstrate a risk assessment problem in an Iranian leading insurance company. Car insurance customers of this company have been analyzed with six different data mining algorithms (C5, Classification and Regression Tree, Neural Networks, Logistic Regression, Bayesian Networks and Support Vector Machines) in two different approaches. One of these approaches is a direct approach in which the target field (risk) is predicted directly with data mining algorithms and then an ensemble model comprised from them. The other one is an indirect approach in which the target field would be divided in five fields, then five different ensemble models is comprised for each new target field. Afterwards the model with the highest confidence predicts the target fields for a test data record. At the end of this article the better results of indirect model would be shown.

Keywords: Assessment, Insurance Industry, Car Insurance, Data Mining, Insurance Risk

1. Introduction

In today's financial industry, credit risk assessment is a critical issue. According to Fair Issac Company, over 75% of mortgage lenders and over 90% of credit card lenders use credit scores when making lending decisions. Higher credit scores results in higher probability of accepting their proposal and lower interest rates [1]. It's designed to predict risk, specifically, the likelihood that applicant will become seriously delinquent on his/her credit obligations in the 24 months after scoring [2].

Financial organizations can differentiate between good and bad customers by credit scoring and credit assessment of their applicant; therefore they can survive in industry and achieve their supreme goals. Credit scoring or credit risk assessment is an important research issue in the financial industry. The major challenge of credit scoring is to recruit the profitable customers by predicting the bankrupts [3].

Financial organizations apply history of applicant for credit scoring and risk assessment of them. Customer's history consists of payment history, length of credit history and types of services in use [4]. Financial organizations assess the credit risk of their customers with a variety of techniques. Since Insurance markets are special cases of markets for contingent claims [5], risk assessment is of great necessity in this industry and several tools and methods are used to do this job. For instance, In 1999 Gourieroux used an econometric approach for risk classification in insurance. He used various information like data on conditional characteristics or data on claim histories or endogenous insurance demand by the agent, to discuss about some questions related to risk classification [6]. In another study, based on the theory of adverse selection and the theory of moral hazard and by using the car insurance data of the individuals, Richaudeau gave some estimation about the risk of the insured by means of a two-step maximum-likelihood method [7]. One of the best and most applied methods for this aim is data mining (DM). DM methods predict the customer behavior, so they assess

customer's risk by their historical data [8]. A lot of research has been done in credit risk assessment of loan customers of financial organizations with DM methods. Some of these researches are described in this section.

The typical method for assessing credit risk of customer is filling a form, and score to each item in this form. This form is analyzed and scored by experts. Finally the summation of scores in this table would show applicant's credit. But these score tables cannot easily change when condition changes. So Li et al. proposed a method called IGCSM. This method applies ID3 decision tree for choosing items. In addition, it uses node's information gain for estimating item's score. This method demonstrates higher correctness when condition changes [9]. Besides, Yin and Lu in 2010 proposed applying clustering and classification methods to reduce items and indexes which are used for risk assessment [10].

In 2009, Hsu and Hung compared SVM and multiple discriminate analyses (MDA) for a credit rating problem. Their study shows that SVM model has better results [11]. Moreover in 2010 a summary research has been done on risk assessment by three DM methods. These methods were logistic regression, classification and regression tree (CART) and neural networks, which are the most common methods in classification. The results show that neural network model is a little bit better than the other methods [12]. Hammer et al. in 2012 applied logistic regression, SVM and logical analysis of data (LAD) for assessing applicant's credit scoring. Their study shows better result of LAD compared with SVM and logistic regression [13]. Furthermore a fuzzy probabilistic rough set model was applied for credit scoring evaluation by Capotorti and Barbanera [14]. Another risk assessment is done by Feki, Ishak and Feki. They applied Bayesian and multi-class SVM. For their problem, SVM model shows superior results compared with Bayesian model [15]. Instance-based models are other DM methods which are applied for credit risk evaluation [16].

SVM is one of the most reputable DM methods used for credit risk assessment in commercial organizations. SVM has weakness in classifying when training data has much noise and redundancy which consequent in low classification accuracy. Feng et al. proposed applying SVM based on principal component analysis (PCA-SVM). This new approach improved processing speed and classification accuracy. At last they compared their approach by SVM and BP neural networks in a risk assessment problem and their approach eventuated in better results [17]. Zhao Min used a similar approach. He applied fuzzy SMV classification model based on PCA (PCA-FSVM). This approach showed better performance and classification accuracy compared with SVM and BP neural networks [18]. Multi-class support vector machines (MSVMs) is another form of SVMs which was used for credit scoring by Kyoung-jae Kim and Hyunchul Ahn [19]. In 2008 Zhang et al. introduced a hybrid credit scoring model (HCSM) by incorporating advantages of genetic algorithms and SVMs. Results demonstrated a better accuracy of this method compared with SVMs, genetic programming, decision trees, logistic regression and back propagation neural network [20]. In 2012 Chen et al. proposed a new hybrid data mining technique for credit risk assessment. They used a two-stage method. In the first stage a clustering method was applied. Isolated samples in the model were deleted and inconsistent samples were relabeled. Then in the second stage (classification stage), SVM algorithm has been applied. These two stages eventuated in classifying samples in three or four classes [21]. As it can be seen, SVM is a suitable algorithm for risk assessment.

In 2009 a hybrid method was proposed by Chen, Ma and Ma. They applied SVM based on CART and MARS to select input features and also used grid search to optimize model parameters [22]. Xiang Liu and Xiaomin Zhu applied a new two-step approach. In the first step they used information gain method to screen the alternative indicators which had greater impacts on prediction. Then in the second step they applied logistic regression modeling method to predict the credit risk of applicant. This method was named ICRES (Individual Credit Risk Evaluation System) [23]. Chiu et al. proposed a new method in DM which is applied for credit risk assessment. They synthesis rough sets and decision tree algorithm. This new method retains the internal features of the original data, speeds up the process of access to knowledge, improves the classification accuracy rate, enhances the interpretability of the rules, and achieves satisfactory results [24].

There are a lot of different artificial intelligence and statistical methods for assessing credit risk of applicant, but which ones are better? Ensemble methods can improve the performance of these problems. Wang et al. studied three ensemble methods (bagging, boosting and stacking) with four statistical and artificial intelligence techniques (logistic regression, decision tree, artificial neural networks and SVM). They demonstrated better results when using ensemble methods for risk assessment [25]. Another example of using ensemble methods for risk assessment was RSB-SVM, a new hybrid ensemble approach which was applied by Wang and Ma. This method used bagging and random subspace based on SVM as base learner [26]. Combining ensemble methods can help us achieve a better credit evaluation. Combining bagging and AdaBoost is one example for this approach. This two-level classifier eventuates in superior results in credit scoring assessment problems, and in general, classification problems [27].

In this paper we are going to apply DM methods to evaluate the risk for Iranian car insurance customers. Our database consists of an attribute which demonstrates the risk of applicants. This attribute has five different classes, which are high risk applicants, risky applicants, moderate risk applicants, low risk applicants and safe applicants. Because of the conditions of this problem, we propose using a special ensemble approach.

2. Research Method

In this research we applied CRISP-DM (CRoss Industry Standard Process for Data Mining) as our methodology.

This standard methodology consists of six steps. These six steps are shown in figure 1 and briefly introduced in continue.

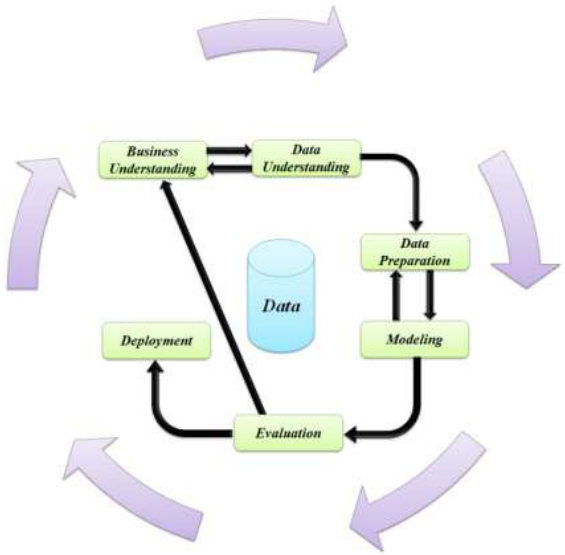

Figure 1. *CRISP-DM methodology.*

CRISP-DM steps are [28]:

Business Understanding: This initial phase focuses on understanding the project objectives and requirements from a business perspective, and then converting this knowledge into a data mining problem definition, and a preliminary project plan designed to achieve the objectives.

Data Understanding: The data understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data, or to detect interesting subsets to form hypotheses for hidden information. There is a close link between Business Understanding and Data Understanding. The formulation of the data mining problem and the project plan requires at least some understanding of the available data.

Data Preparation: The data preparation phase covers all activities to construct the final dataset (data that will be fed into the modeling tool(s)) from the initial raw data. Data preparation tasks are likely to be performed multiple times, and not in any prescribed order. Tasks include table, record, and attribute selection, data cleaning, construction of new attributes, and transformation of data for modeling tools.

Modeling: In this phase, various modeling techniques are selected and applied, and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques require specific data formats. There is a close link between Data Preparation and Modeling. Often, one realizes data problems while modeling or one gets ideas for constructing new data.

Evaluation: At this stage in the project you have built one or more models that appear to have high quality, from a data analysis perspective. Before proceeding to final deployment of the model, it is important to more thoroughly evaluate the model, and review the steps executed to construct the model, to be certain it properly achieves the business objectives. A key objective is to determine if there is some important business issue that has not been sufficiently considered. At the end of this phase, a decision on the use of the data mining results should be reached.

Deployment: Creation of the model is generally not the end of the project. Usually, the knowledge gained will need to be organized and presented in a way that the customer can use it. Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process. In many cases it will be the user, not the data analyst, who will carry out the deployment steps. In any case, it is important to understand up front what actions will need to be carried out in order to actually make use of the created models.

In findings section this methodology would be applied.

Moreover we want to assess the risk of car insurance customers in two different approaches and then compare them. One of these approaches is a direct approach in which we would develop an ensemble model for the target field from five different algorithms (C5, C&RT, Neural Networks, Logistic Regression, and Bayesian Networks). And in the other approach we would develop five ensemble models for each value of the main target field (risk field transformed to five fields which consist of very safe, safe, neutral, risky, and very risky) from six different algorithms (C5, C&RT, Neural Networks, Logistic Regression, Bayesian Networks, and SVM).

3. Findings

In this section we would describe our analysis for a car insurance database (a database for an Iranian leading insurance company). As we explained in the methodology section, CRISP-DM approach would be applied for this research.

3.1. Business Understanding

In the literature section we explained this research's business understanding in detail. As we described, risk assessment is a critical issue in financial industry, especially in insurance industry. A simple method for credit scoring and risk assessment is filling a form and scoring each item for a customer and then summing up the scores. However, this method is not practical for a huge data base, and machine learning algorithms could simplify this time consuming procedure. Moreover human faults can weaken the results. Therefore data mining methods can be used for assessing customers risk and classifying them in different risk groups. In this research, we want to classify one of the Iranian leading insurance company's customers in five

different classes (high risk, risky, neutral, safe, very safe), and for this purpose we would apply data mining techniques.

3.2. Data Understanding

This article's car insurance data set consists of 27754 data records with three field types (personal, car characteristic and behavioral). Car characteristic and behavioral fields are shown in figure 2 and personal fields are hidden for privacy preserving. Personal fields are customer characteristics like sex, age, name and so on. Car characteristic fields are Car Type, brand and model of the car (Car Name), color of the car (Color) and its application (Car Application), estimated market value of the car (Car Value), its production year (Prod. Year) and also insurance premium(Premium). Furthermore, behavioral fields are Discount (yes for customers who gain discount and no for others), Payment Method (cash or installments), Revival (number of years which customers renew him/her insurance policy), Number of Accidents, and Compensation. This data set is select from the complete data records from original data base, as a result its quality is satisfying (there is no blank and missing data). A section of this data set is shown in figure 2.

Car Type	Car Name	Color	Car Application	Car Value	Prod. Year	Premium	Discount	revive	Payment Method	of Accider	Compensation
Passenger Car	Renault- Megan	Silver	Personal	320,000,000	1388	7,324,000	No	3	Instalment	2	3,800,000
Passenger Car	Renault- Megan	White	Personal	320,000,000	2008	5,338,000	No	0	Cash	1	4,500,000

Figure 2. *car characteristic and behavioral fields for this article's case study.*

3.3. Data Preparation

When we surveyed data set, some problems were seen. The most important problem in this data set is data entry faults, for example a number of data records have only brand or model in car name (e.g. only Renault or only Megan) which would comprise lots of problems in analysis. Furthermore risk assessment is not applied for all of them. Therefore we integrated data set and assessed risk for all of them. The next preparation issue is creating train and test datasets. For this reason we randomly select 7000 data records for testing our risk assessment model and 20754 data records for training it. Also we want to test two different approaches for risk assessment, one direct approach and compare it to a multi target approach. Therefore creating multi target data from original ones is the other effort in this step (risk field transformed to five fields which consist of very safe, safe, neutral, risky, and very risky).

3.4. Modeling

In this step we develop two models. One of them (as described) is an ensemble model from C5, C&RT, Neural Networks, Logistic Regression and Bayesian Networks algorithms and the other one is a model consisting of five ensemble model for each risk class (very safe, safe, neutral, risky, and very risky) from C5, C&RT, Neural Networks, Logistic Regression, Bayesian Networks, and SVM algorithms. In this section we would describe C5, C&RT, Neural Networks, Logistic Regression, Bayesian Networks, and SVM algorithms. Five ensemble models must be also compounded. Therefore the highest confidence ensemble model (from five ensemble models for five target field classes) would predict the class of a data record. In the evaluation step the results of two models would be illustrated.

3.4.1. C5

C4.5 builds decision trees from a set of training data in the same way as ID3, using the concept of information entropy. At each node of the tree, C4.5 chooses the attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. The splitting criterion is the normalized information gain (difference in entropy). The attribute with the highest normalized information gain is chosen to make the decision. This procedure recursively continue until stopping criteria met [29]. Quinlan went on to create C5.0. C5.0 offers a number of improvements on C4.5 like better speed, efficient memory usage and smaller decision tree [30].

3.4.2. Cart

Classification and regression tree is a classification method which uses historical data to construct decision trees. CARTs are machine-learning methods for constructing prediction models from data. The models are obtained by recursively partitioning the data space and fitting a simple prediction model within each partition. As a result, the partitioning can be represented graphically as a decision tree [31], [32].

3.4.3. SVM

Support vector machines (SVMs), a method for the classification of both linear and nonlinear data. SVM is an algorithm that uses a nonlinear mapping to transform the original training data into a higher dimension. SVMs are based on the concept of decision planes that define decision boundaries. A decision plane is one that separates between a set of objects having different class memberships. With an appropriate nonlinear mapping to a sufficiently high dimension, data from two classes can always be separated by a hyper plane. The SVM finds this hyper plane using support vectors and margins [8], [17].

3.4.4. Bayesian Networks

Bayesian networks belong to the family of probabilistic graphical models. These graphical structures are used to represent knowledge about an uncertain domain. In particular, each node in the graph represents a random

variable, while the edges between the nodes represent probabilistic dependencies among the corresponding random variables. These conditional dependencies in the graph are often estimated by using known statistical and computational methods [33]. Bayesian classifiers are statistical classifiers. They can predict class membership probabilities such as the probability that a given data belongs to a particular class [8].

3.4.5. Neural Networks

A neural network is an information processing paradigm that is inspired by the way biological nervous systems, such as the brain, process information. Roughly speaking, a neural network is a set of connected input/output units in which each connection has a weight associated with it. Neural networks learn from data, like human. A neural network is configured for a specific application, such as classification, through a learning process. During the learning phase, the network learns by adjusting the weights, to be able to predict the correct class label of the input data [8], [34].

3.4.6. Logistic Regression

Logistic regression is part of a category of statistical models called generalized linear models. Logistic regression allows predicting a discrete outcome, such as class membership, from a set of variables that may be continuous, discrete, dichotomous, or a mix of any of these [35]. The logistic model formula computes the probability of the selected response as a function of the values of the predictor variables [36].

3.4.7. Ensemble Model

An ensemble combines a series of k learned models (or base classifiers), with the aim of creating an improved composite classification model. An ensemble tends to be more accurate than its base classifiers. This article's ensemble classifier collects the class label predictions returned from the base classifiers and outputs the class in majority. The base classifiers may make mistakes, but the ensemble will misclassify a data record only if over half of the base classifiers are in error. Ensembles yield better results when there is significant diversity among the models. That is, ideally, there is little correlation among classifiers [8].

3.5. Evaluation

After developing models, they must be analyzed with the test data set. Besides, a base scenario must be considered in which all data records are predicted as neutral (with no risk assessment algorithms, all customers are predicted as neutral). The results have shown 48.67% (direct model) and 62.96% (indirect model) accuracy, which must be compared with 13.19% accuracy for the base scenario. Accuracy itself is not the only method for comparing these models. In Table 1, the coincidence matrix (rows show actual risk of the test data set and columns show the predicted risk for them) for direct model and in Table 2, the coincidence matrix for indirect model, are shown. As it can be seen, indirect model has a better prediction ability for identifying very risky customers which are the most important target for a risk assessment model. Besides, if a very risky customer is predicted as a very safe customer, insurance company would lose more money. So with respect to different misclassification costs, the direct model costs 87.84% of the base scenario costs and indirect model costs 51.39% of the base scenario costs. Although indirect model shows better results, both models are accepted.

Table 1. *coincidence matrix for direct model.*

Direct	Very Safe	Safe	Neutral	Risky	Very Risky
Very Safe	382	35	24	41	14
Safe	256	210	154	144	64
Neutral	183	80	393	209	58
Risky	125	58	244	1867	25
Very Risky	798	233	490	358	555

Table 2. *coincidence matrix for indirect model.*

Indirect	Very Safe	Safe	Neutral	Risky	Very Risky
Very Safe	302	62	16	54	62
Safe	174	222	87	199	146
Neutral	93	96	305	272	157
Risky	79	41	70	1962	167
Very Risky	181	64	115	458	1616

3.6. Deployment

This step is beyond this research and is dedicated to the insurance company.

4. Conclusion

In today competitive world, customers can select products and services. This can make competition difficult, which leads companies to more productive approaches and procedures. One of the most competitive services is insurance. Insurance companies fight for attracting customers; however some of customers are not preferable and are risky. Therefore insurance companies must predict their customers' risk. This problem for an Iranian leading insurance company has been studied and two different models have been developed for it, in this article. These models are data mining models, which are developed with CRISP-DM methodology. One of these models was following a direct approach. This model is an ensemble model which has been developed with five different data mining algorithms (C5, C&RT, Neural Networks, Logistic Regression, and Bayesian Networks). The other model was pursuing an indirect approach. This model consists of five ensemble models for each class of target field (risk), in which the model with the highest confidence predicts risk label for car insurance customers of Iranian leading insurance company. These models were ensemble models which have been developed with five different data mining algorithms (C5, C&RT, Neural Networks, Logistic Regression, Bayesian Networks and SVM). The results showed better results of the indirect model.

References

[1] Credit Scoring. [Internet]. [cited 2012 Nov 25]. Available from: http://epic.org/privacy/creditscoring/.

[2] What is a credit score? [Internet]. [cited 2012 Nov 25]. Available from: http://www.bankrate.com/finance/credit-cards/what-is-a-credit-score.aspx.

[3] Marikkannu P, Shanmugapriya K. Classification of customer credit data for intelligent credit scoring system using fuzzy set and MC2 — Domain driven approach. Electronics Computer Technology (ICECT). 2011 410-414.

[4] What's in my FICO Score. [Internet]. [cited 2012 Dec 25]. Available from: http://www.myfico.com/crediteducation/whatsinyourscore.aspx.

[5] Koeniger W. Labor Income Risk and Car Insurance in the UK. The GENEVA Papers on Risk and Insurance Theory. 2004;29(1):55-74.

[6] Gourieroux C. The Econometrics of Risk Classification in Insurance. The GENEVA Papers on Risk and Insurance Theory. 1999;24:119-137.

[7] Richaudeau D. Automobile Insurance Contracts and Risk of Accident: An Empirical Test Using French Individual Data. The GENEVA Papers on Risk and Insurance Theory. 1999;24(1):97-114.

[8] Han J, Kamber M, Pei J. Data Mining Concepts and Techniques. Waltham: Morgan Kaufmann; 2012.

[9] Li F, Xu J, Dou ZT, Huang YL. Data mining-based credit evaluation for users of credit card. Machine Learning and Cybernetics. 2004 2586- 2591.

[10] Yin Q, Lu K. Data mining based reduction on credit evaluation index of bank personal customer. Future Information Technology and Management Engineering (FITME). 2010 570-573.

[11] Hsu CF, Hung HF. Classification Methods of Credit Rating - A Comparative Analysis on SVM, MDA and RST. Computational Intelligence and Software Engineering. 2009 1-4.

[12] Wah YB, Ibrahim IR. Using data mining predictive models to classify credit card applicants. Advanced Information Management and Service (IMS). 2010 394- 398.

[13] Hammer PL, Kogan A, Lejeune MA. A logical analysis of banks' financial strength ratings. Expert Systems with Applications. 2012 7808-7821.

[14] Capotorti A, Barbanera E. Credit scoring analysis using a fuzzy probabilistic rough set model. Computational Statistics and Data Analysis. 2012 981–994.

[15] Feki A, Ishak AB, Feki S. Feature selection using Bayesian and multiclass Support Vector Machines approaches: Application to bank risk prediction. Expert Systems with Applications. 2012 3087–3099.

[16] García V, Marqués AI, Sánchez JS. On the use of data filtering techniques for credit risk prediction with instance-based models. Expert Systems with Applications. 2012 13267–13276.

[17] Feng W, Zhao Y, Deng J. Application of SVM Based on Principal Component Analysis to Credit Risk Assessment in Commercial Banks. Intelligent Systems. 2009 49-52.

[18] Min Z. Credit Risk Assessment Based on Fuzzy SVM and Principal Component Analysis. Web Information Systems and Mining. 2009 125-127.

[19] Kim Kj, HyunchulAhn. A corporate credit rating model using multi-class support vector machines with anordinal pairwise partitioning approach. Computers & Operations Research. 2012 1800–1811.

[20] Zhang D, Hifi M, Chen Q, Ye W. A Hybrid Credit Scoring Model Based on Genetic Programming and Support Vector Machines. Natural Computation. 2008 8-12.

[21] Chen W, Xiang G, Liu Y, Wang K. Credit risk Evaluation by hybrid data mining technique. Systems Engineering Procedia. 2012 194 – 200.

[22] Chen W, Ma C, Ma L. Mining the customer credit using hybrid support vector machine technique. Expert Systems with Applications. 2009 7611–7616.

[23] Liu X, Zhu X. Study on the Evaluation System of Individual Credit Risk in commercial banks based on data mining. Communication Systems, Networks and Applications (ICCSNA). 2010 308-311.

[24] Chiu JY, Yan Y, Xuedong G, Chen RC. A New Method for Estimating Bank Credit Risk. Technologies and Applications of Artificial Intelligence (TAAI). 2010 503- 507.

[25] Wang G, Hao J, Mab J, Jiang H. A comparative assessment of ensemble learning for credit scoring. Expert Systems with Applications. 2011 223-230.

[26] Wang G, Ma J. A hybrid ensemble approach for enterprise credit risk assessment based on Support Vector Machine. Expert Systems with Applications. 2012 5325–5331.

[27] Marqués AI, García V, Sánchez JS. Two-level classifier ensembles for credit risk assessment. Expert Systems with Applications. 2012 10916–10922.

[28] Wirth R, Hipp J. CRISP-DM: Towards a Standard Process Model for Data Mining. In: 4th International Conference on the Practical Application of Knowledge Discovery and Data Mining (KDD '96); 2000; Manchester, UK. p. 29-39.

[29] Quinlan JR. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers; 1993.

[30] Quinlan JR. Improved use of continuous attributes in c4.5. Journal of Artificial Intelligence Research. 1996 77-90.

[31] Loh WY. Classification and regression trees. John Wiley & Sons. 2011 14-23.

[32] Timofeev R. Classification and Regression Trees (CART) Theory and Applications. Berlin: CASE - Center of Applied Statistics and Economics, Humboldt University; 2004.

[33] Ruggeri F, Faltin F, Kenett R. Bayesian Networks. Encyclopedia of Statistics in Quality & Reliability. 2007 1-6.

[34] Stergiou C, Siganos D. NEURAL NETWORKS. [Internet]. [cited 2012 Nov 27]. Available from: http://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/cs11/report.html#What is a Neural Network.

[35] Logistic Regression. [Internet]. [cited 2012 Nov 27]. Available from: http://userwww.sfsu.edu/efc/classes/biol710/logistic/logisticreg.htm.

[36] Logistic Regression. [Internet]. [cited 2012 Nov 28]. Available from: http://www.dtreg.com/logistic.htm.