Indian Journal of Science and Technology, Vol 9(18), DOI: 10.17485/ijst/2016/v9i18/93767, May 2016
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Identifying the Effective Factors in the Profit and Loss of Vehicle Third Party Insurance for Insurance Companies via Data Mining Classification Algorithms

Karamizadeh Faramarz[1], Zolfagharifar Seyed Ahad[2*] and Dastghaibyfard Gholamhosseyn[1]

[1]Department of Electrical and Computer Engineering, Shiraz University, Shiraz, Iran; f.karamizadeh@gmail.com; dstghaib@shiraz.ac.ir

[2]Islamic Azad University Science and Research, Kohgiluyeh and Boyer Ahmad Branch, Iran; sir.zolfaghari@gmail.com

Abstract

Background: Insurance companies for surviving and keeping the market, always emphasize on profitability and reduction of their losses. **Methods:** Investigation of the Car Insurance Information shows that main factors which effect on profit or loss of insurance companies include: type of vehicle usage, license, type of license and compliance or lack of compliance with the vehicle, the amount of the premium, rate of the commitment, the car quality of the car companies, the age of the driver, driver education, the mismatch premiums with the insurance case, the delay in the renewal of insurance policies. **Findings:** In this paper, using data mining information of 2011year of third party insurance in Iran insurance companies in Kohgiluyeh and Boyer Ahmad province are studied. The results showed that using the classification algorithm with over 91% accuracy and decision trees with over 96% accuracy able to provide a model to identify effecting factors and determine their impact on the profit and loss of vehicle third party insurance. **Applications/Improvements:** Comparing results indicated that, the decision tree Wj48 with very high accuracy will able to detect and predict the occurrence of the damage of an insurance, properly. After that two other algorithms as well as with high accuracy and the same have this ability.

Keywords: Classification Algorithm, Data Mining, Insurance Companies, Profit and Loss, Third Party Insurance

1. Introduction

Insurance companies for survive and keep the market always emphasize on profitability and reduce their losses. Factors such as marketing, customer loyalty, premium rates, promotions, fraud can attract or repel customers that in the profit and loss have direct or indirect impact[1].

One way to reduce the losses review and analyze the damage data using data mining[2]. The purpose of data mining detecting trends or patterns to make better decisions that this objective by employing statistical methods such as logistic regression, clustering as well as using data analysis methods realize[3].

Factors that caused the damage and ultimately increase the losses an insurance company can be driving culture, have driving license, type of license, and compliance or non-compliance with the vehicle, the structure of urban roads and public transit, fraud, weather statues, car quality of car makers, driver age, driver education, the mismatch of premiums with insurance case[4], holidays, travel and many others, and ultimately cause increase losses an insurance company. With collective these factors in any event and other information with using data mining (with supervisor or without supervisor) can discover the hidden rules in the data achieved[5]. In the methods of with supervisor with using the compensation sample

*Author for correspondence

built a model that based on determined damage or non-damage of a new sample[6].

2. Overview on the Done Work

With the help of logistic regression algorithm discovered the fraud. Wilson, in this study could with 70.4% accuracy predict samples of fraud. He was able to show that with this model with an accuracy of 81.6% real claims distinguish from unreal samples.

To detect fraud in the car insurance. The combination of propagation neural network algorithms and simple Bayes and decision tree C4.5 led to the discovery of insurance fraud in their research[7].

With using data mining provide a framework to predict the car hull insurance customer damage that during it, the predicted risk rate of customers and customers accordingly it classified[8]. Evaluation methods used in this study to two categories of internal evaluation and external divided. The correlation coefficient was calculated for both them is 82% which shows that the results of both methods as desired with each match together.

3. Data and the Research Implementation Environment

In this study, at first information of export insurance and third-party damage in 2011 year of in Kohgiluyeh and Boyer Ahmad province (about 20 thousand records that about 1500 records had damage) which contains 179 export data fields, was collected. Then, 137 field of them which did not affect, removed and the finally, final effective fields were reduced to 42 fields (Tables 1 and 2).

Data mining operations was done by Rapid miner software and to optimize the response and quality of the results in the areas used Minitab software and Clementine12 (Table 3).

3.1 Steps of Research Implementation

The steps includes of software selection, data entry, fields of export data sets and loss, reduce the dimension and cleaning up data, address to the data lost, outlier data discovery and data transmission to the data mining environment and implementation of algorithms.

Table 1. Data selected fields of insurance damage

Data selected fields
Amount of damage
Date of accident creation
The first insurer seen loss
The number of the injured suffered
The number of deceased suffered

Table 2. Data selected fields of insurance export

Name of field	*Rowe*	*Name of field*	*Rowe*	*Name of field*	*Rowe*
start date	29	Surplus commitment	15	Month	1
Date of issue	30	Physical commitment	16	Year	2
Organization Name	31	Financial commitment	17	The main exporting agency code	3
Insurance numbers	32	Insurance of last year	18	Group discounts	4
Employee	33	Type of insurance	19	Discounts no damage	5
Issued by branch	34	Type of plate	20	Type of Document1	6
Government	35	Description the used case	21	forfeiture delay	7
Representative Place of Issue	36	Capacity	22	Add code of premium rates	8
Damage?	37	Number of cylinders	23	Received premium	9
The amount of damage	38	Year of construction	24	Effects of Article 92	10
Date of accident creation	39	System	25	Taxation	11
The first insurer seen loss	40	Type of vehicle	26	Premium of passengers	12
The number of the injured suffered	41	Term of insurance	27	Excess Premium	13
The number of deceased suffered	42	Expiration date	28	law third-party premium	14

Table 3. Statistics export third party insurance of Kohgiluyeh and Boyer Ahmad province in 2011 year

Methods	*Total of records*	*Number of effective records*	*fields*		*Applications software*
			Ineffective	*Effective*	
Data mining	20000	1500	137	42	Rapid miner Minitab, Clementine 12

3.2 Deal to the Lost Data

In the initial phase with sorting all the features that available sort in the Microsoft Excel software to discover the lost values and through the other characteristics of the record guessed the lost values. In some cases due to the large number of missing features, we have to remove the full record.

3.3 Outlier Data Discovery

To detect outlier data was used the chart of box plot from Minitab15 software. Values greater than, Q3 + [(Q3 - Q1) X 1.5] and less than Q1 - [(Q3 - Q1) X 1.5], are outliers data.

3.4 Selection Operation of more Effective Properties

In dealing with some of the algorithms that by increasing the number of features have more complex. In general, feature selection for use in the classification algorithm is the strategic technique. Feature selection algorithms used include PCA[11], SVM Weighting, Weighting Deviation, Weighting Correlation that, according to Table 4 are selected fields.

4. Evaluation Criteria of Classification Algorithms

A) Confusion Matrix

This matrix shows the classification algorithm performance according to the input data sets divided type of issue categories of classification. In this matrix, concepts of TN[22], FP[3], FN[4], TP[5] is as follows[9] (Table 5):

TN: Indicate the number of records that their actual category were negative and the algorithm properly detected their category was negative.

FP: Indicate the number of records that their actual category were negative and the algorithm wrong detected their category was positive.

Table 4. Results of field community with the highest weight in the different algorithms

Name of field	Type of field
Surplus commitment-Physical commitment-Financial commitment-Type of plate-Capacity- Number of cylinders-Year of construction- Term of insurance-The number of the injured suffered- The number of deceased suffered	Integer
Description the used case -System- Type of vehicle-The first insurer seen loss	Polynomial
forfeiture delay- Add code of premium rates-Received premium-Taxation- Premium of passengers-law third-party premium- The amount of damage	Real
Insurance of last year-Employee- Issued by branch	Binominal

Table 5. Confusion Matrix

(Actual Records)	(Predicted records)	
Category+	Category-	
FP	TN	Category-
TP	FN	Category+

FN: Indicate the number of records that their actual category were positive and the algorithm wrong detected their category was negative.

TP: Indicate the number of records that their actual category were positive and the algorithm properly detected their category was positive.

The most important criteria for determining the efficiency of an algorithm classification is classification accuracy[1] criteria. This criteria shows that what percentage of all training records sets are properly classified.

Classification accuracy calculated by the following equation:

$$CA^{1} = \frac{TN + TP}{TN + FN + TP + FP} \quad (1)$$

B) AUC[21] Criteria

This criteria for determining the efficiency of a classifier is very effective. This criteria represents the area under the ROC[3] curve. The number AUC due to a larger classifier,

Principle Component Analysis[1]
True Negative[2]
False Positive[3]
False Negative[4]
True Positive[5]
Area Under Curve[2]
Receiver Operating Characteristic[3]

classifier final performance is more favorable. In the ROC, the rate of positive category correct diagnosis on the axis Y and the rate of negative category misdiagnosis plotted on the X axis. If any axis is ranges between 0 and 1, the best spot on this criteria (0,1) and points (0,0) is positive classifier and false alarm never produced.

4.1 Evaluation Methods of Classification Algorithms[10]

In learning methods with supervisor that classification is one of these methods, two important data collection the name of the training data and test data are available.

4.1.1 Holdout Method

In this method, the original data collection divided in two part of training and experimental. Classification model by training data set made and will be assessed by experimental data set[11,12]. The data collection division ratio depends on the analyst recognition, but two of the better-known ratio in this method are as follows 50–50 and or two-thirds for training and a one-third for experimental and evaluation.

4.1.2 Random Subsampling Method

If the Holdout method run several times and from the results averaging we have got a more reliable method which Random Subsampling is called.

4.1.3 Cross-Validation Method

If the Random subsampling method any of the records, in equal numbers for learning and only use once for evaluation, we have taken a more intelligent way. This method in the scientific literature Cross-Validation called.

4.1.4 Bootstrap Method

If a record given more than once in the model learning operation Bootstrap method have been adopted. In this method, training records to do the model learning process from the initial data collections, will be selected as sampling with replacement and records did not select use for the evaluation.

5. Methodology

In the implemented classification algorithm every three Holdout, k fold Validation, Bootstrap used, and the results were compared. In the Holdout method which in the software called Split Validation, the standard ratio is 70% of primary data collection for training and 30 percent used for testing. For k fold Validation, k value considered equal to 10 which is the standard value. In the Bootstrap also the data set divided value considered 10 episodes. The local random seed value also equal to 1234567890, for all models, the software used it unless in the specific model, do not use it or change the value cause the algorithm performance improvement, that is indicating.

In this study, 8 classification algorithm include KNN, Naïve bayes, Neural Network, SVM Linear, Meta Decision Tree, Wj48, Random Forest and logistic regression was used, the number of 3 algorithm were Decision Tree. The classifier algorithms results of other trees compared with together and respectively identified the best models. Also, three tree classification algorithm compared with each other and best results being damage of a record determines, after consultation with experts and insurance experts mined from every tree and respectively ultimate accuracy of the algorithm has been announced.

6. Classification Algorithms

The result of comparing comes in Tables 6 and 7 that shows, logistic regression algorithm and SVM algorithm with 98.54% ability to identify correctly predicted the loss of a record. Then simple Bayes algorithm with a 96.09%, and then neural network algorithm and KNN nearest neighbor by K = 11 with about the same accuracy of ability to detect and predict the loss of an insurance case.

For all classification algorithms drawn the AUC diagram, here the diagram of the KNN algorithm came for example (Figure 1).

Implementation of decision trees algorithms containing forms of the trees which are as follows (Figures 2, 3).

In the tree algorithm random forest number of 20 trees were produced which in Figure 3 shown one example of

Table 6. Decision Tree classification algorithm comparison

The best evaluation model	The accuracy of detection%	Name of algorithm
10 Fold Validation	99.52	Wj48
Split Validation	96.72	Random Forest
Split Validation	96.64	Meta Decision Tree

Table 7. Classification algorithm comparison

The best evaluation model	The accuracy of detection%	Name of algorithm
Split Validation	98.54	Logistic regression
Split Validation	98.54	Linear SVM
Split Validation	96.09	Naïve Bayes
Split Validation	91.25	Neural Network
Split Validation	91.23	KNN

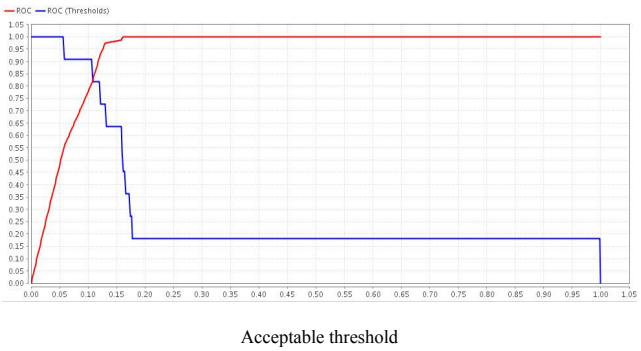

Figure 1. AUC Diagram KNN algorithm.

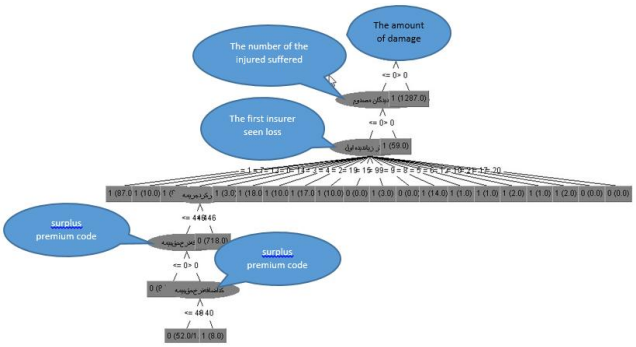

Figure 2. Tree Diagram Wj48 algorithm.

it. Analytical results from these algorithms is the result community of 20 trees.

7. Conclusion

Compare the results show that, the decision tree Wj48 with very high accuracy will able to detect and predict the occurrence of the damage of an insurance. After that two other algorithms as well as with high accuracy and the same have this ability. Best intuitive results of this trees after check the insurance experts includes if forfeiture delay of a person is less than 13,000 rials and or without forfeiture, with 81% probability will not load the damage. According to the

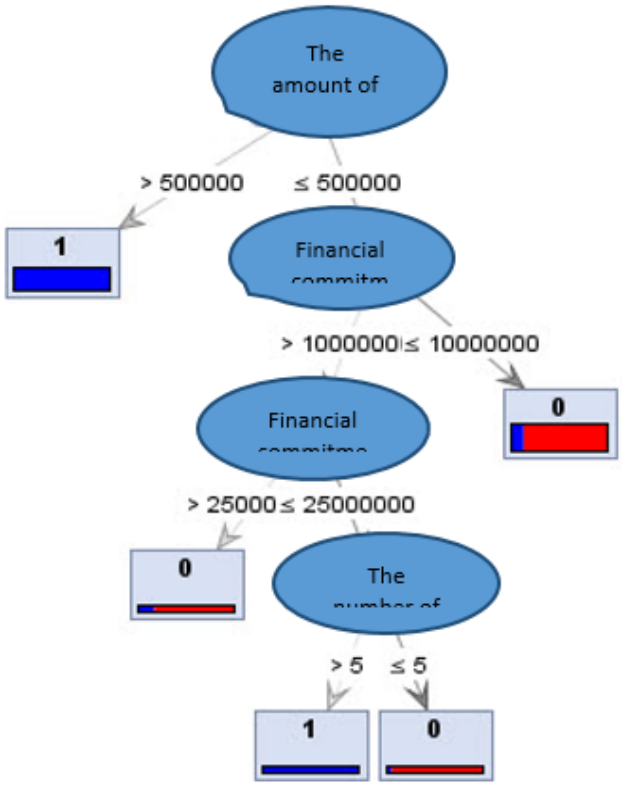

Figure 3. A tree sample produce by the Random Forest algorithm.

accuracy matrix, rule of number one is 81%. That occurred in 81% of cases, this law will be correctness. Insurance Expert opinion: This indicates that a person has carefully kept the expiry date of his insurance, to the extent that just with one day of delay, to renew their insurance.

If forfeiture delay of a person is 13000 rials (more than one day) and addition code of insurance rate is more than 30, probably 81%, will bring damage. Addition code of rate awarded to individuals who from their previous insurance place caused the damage and cutting their insurance slowed. If a person do not have forfeiture delay or have less than a day delay and the amount of discounts no damage is more than 78% of law premium-party:

A) If he used car is taxi then with probability 90% will not have damage.
(Confusion matrix law a3). Due to the small number of positive category toward the whole record, so this law is not invoked.

B) However, if the usage case is inter-city taxi, with probability of 30% will bring damage. (Confusion matrix law b3). Due to the small number of positive category toward the whole record, so this law is not invoked.

C) If the vehicle is draft and in April the insurance is renewed, with probability of 60% will not bring damage. (Confusion matrix law c3).
Due to the small number of positive category toward the whole record, so this law is not invoked.

D) If the other months of the year renew 97% will bring damage. (Confusion matrix law d3).
Due to the small number of positive category toward the whole record, so this law is not invoked.

E) If the usage case is administrative, it will not bring 100% damage. (Confusion matrix law e3).
Due to the small number of positive category toward the whole record, so this law is not invoked.

F) If the usage case is between-city taxi (desert fare) with probability of 75% will result in damage. (Confusion matrix law f3).
Due to the small number of positive category toward the whole record, so this law is not invoked.

G) If the usage case is person or personal hire, it will not bring 89% damage. (Confusion matrix law j3)

Insurance, third party insurance obtained from the following formula: law third-party premium +surplus premium + premium of passengers + taxation + effects of article 29+ add rate + late penalty - forfeiture delay -discounts no damage - group discounts.

Meanwhile, law third-party premium is the largest amount of premium and forfeiture delay as well as can be a large amount of the premium which depends on the number of days of delay renew the insurance. According to the formula, person who have discounts no damage equivalent to 78% of law third-party premium then pays a small premium. But the risk of cars according to their usage are different. Because the agencies are not constantly moving and are kind of private cars, which use a dual purpose to take less risk.

In contrast, inter-city taxi due to traffic more frequent are at-risk and between-city taxis due to the long distance to travel and more quickly as well as natural hazards of city roads outside have more risks. The trolleys are in addition to the above conditions the lack of vision of the driver in the cabin mirror is also added. But on the renewal of insurance in April does not have idea however, because time of all insurance is one-year. Meanwhile, official vehicles as well as have low risk because of low traffic and generally have experienced drivers and cars are also healthier. Based on Insurance expert opinion those who have personal vehicles even if they do job usage, they will be more cautious.

Also, must be considered the taxi cars depreciate. Thus, the mere existence of discounts is not an inhibitor of risk. Due to the damage between 40 to 100 percent of taxis and trolleys and despite the mere existence of discounts in payments premium can be paid to the fact that low premium cause indiscretion. Prove this subjects with pointing to discount of more than 78 percent of the premium is justified because the algorithm on less discounts do not tell i.e. when premium is higher, damage have occurred lower. It should be noted that, in the calculation of discounts, actually discounts over than 70% are not calculated and just between the years of discount and the its percentage stated that this issue could be the cause of insured discouraged and, not far away from danger.

As Figure 4 shows, records with discount under 2 million rials more to have been damaged. It represents the power of promotional discounts of no damage at premium and according to that discounts of no damage has directly relationship with insured history, it is clear that this amount of the discount for the first year and second of insured which is effective but in the coming years, the insured is expecting more discounts. Also, the graph shows that the higher the discount, the more discount observe more caution and due to the loss of discounts in case of accident.

About the law "if the car is draft and in April did not renew the insurance with 60% probability ill not be damage" could be provided justified that such people are responsible and at the beginning of each year with starting with the labor market take action to the insurance their own car. But others, when entering the peak time of job and forced conditions take action to the e insurance. Paykan and Pride cars if premium are less than 2,700,000 rials with 40% probability cause damage. (Confusion matrix law 4).

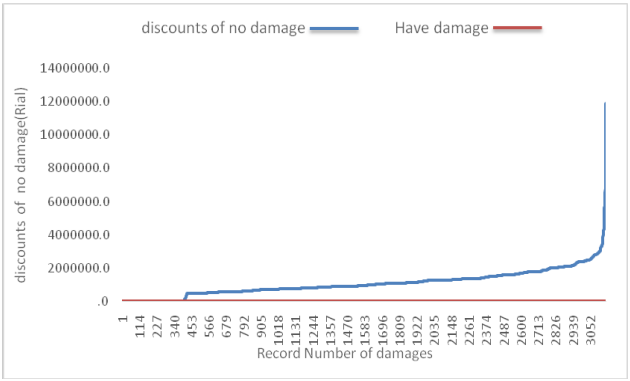

Figure 4. Chart of discounts of no damage to the damage.

This subject could be true. Because cars cheap, low-cost maintenance and spare parts are cheaper. Also this cars are less secure than the more expensive vehicles. Therefore, if a person riding on a small secure car has more risk and if the person knows providing damage could be compensated with a small fee, is less circumspect. Due to the small number of positive category toward the whole records, so this law is not invoked. If a person had not forfeiture delay and his car has more than 15 people capacity with 100% probability will not have damage. (Confusion matrix law 5).

Such vehicles usually are vans or buses which because of speed control by traffic and as well as high security and low speeds, they are very low risk. Due to the small number of positive category toward the whole records, so this law is not invoked.

A- Vehicles with the newer year of manufacture (last three years) with 95% probability will not cause damage. (Confusion matrix law 6).

B- cars are older with 95% probability will be cause damage. (Confusion matrix law a6). According to the statistics show the number of manufactured vehicles damage in the last three years than the proportion of manufactured vehicles in the last three years is 40% less, it can say that law is invoked.

Those who have additional obligations compared to the current commitments purchase, are less likely to produce damage. (Confusion matrix, law 7). Those who purchase additional commitments are forward-looking. These people are likely to produce more damage consider and according to this cause must be observed caution. Toll vehicles and road construction, service personnel and ambulances with 100% probability do not damage. (Confusion matrix, law 8) (Table 8).

Table 8. Confusion matrix

Category+							Category-						
law							law						
3c	3b	3a	3a	2	1	FP	3c	3b	3a	2	1	TN	Category-
4	24	24	1	303	373		8	15	6	618	561		
4	7	7	4	7	1061	TP	4	26	0	5	0	FN	Category+

Category+							Category-						
law							law						
5	4	3j	3f	3e	3d	FP	5	4	3j	3f	3e	3d	TN	Category-
0	35	20	0	0	0		6	28	101	2	4	13		
9	39	165	1	15	25	TP	0	65	12	1	0	1	FN	Category+

Category+					Category-					
law					law					
8	7	6b	6a	FP	8	7	6b	6a	TN	Category-
0	6	34	20		2	561	798	435		
9	860	1121	653	TP	0	10	63	30	FN	Category+

8. Offers

- Calculate discounts over than70% to drivers who have been so cautious that could for several years in a row without risk.
- Calculate the more additional premium for cars with poor structure and unsafe and consequently lower premiums for vehicles safer.
- Applies more discounts for agencies, toll and the construction, service of staff, ambulances and private cars and low traffic vehicles. It can determine low traffic of vehicles based on the job of driver. For example, an employee than a contractor is more low traffic.
- Premiums more expensive for inter-city taxi and a factor of more premium than inter-city taxi for between-city taxis.
- Apply discounts and encouragement to those who without delay provide insurance renewal. Currently, there is a penalty for delay if there is encouragement to renew insurance, can be highly effective.
- Apply discounts for vehicles of low risk group such as minibuses and buses.
- Apply discounts for vehicles of zero km and apply a premium lower than for new cars than depreciate cars.
- Reduce additional commitments sector premium for those who have more commitments than the purchase usual commitments as stepped or progressive.
- Customers losses which is more than double fatality or have 3financial damage pay higher premiums and or are insured just 6 months to cover the period of their risk reduced by half.
- Insert the insurer's profile such as age, occupation, education, history of certification, type of certification or state of health in time of insurance for the science future usage of data mining that would lead to a definitive knowledge will be in the field.
- Insert more information about the accident, place of accident and the scene and the injured party's profile and blamed for the future of data mining.

9. References

1. Derrig R, Johnston D, Sprinkel E. Auto Insurance Fraud: Measurements and Efforts to Combat It. Risk Management and Insurance Review. 2006; 9:109–30.
2. Alpaydin E. MA, USA: The MIT Press Cambridge: Introduction to Machine Learning. 2010.
3. Koh C., Geravis G. Fraud Detection Using Data Mining Techniques: Applications In The Motor Insurance Industry. Proceedings of Business And Information. 2010; 7:49–54.
4. Wilson J. An analytical approach to detecting insurance fraud using logistic regression. Journal of Finance and Accountancy. 2003; 1:1–15.
5. Bolton J, Hand J. Statistical fraud detection: a review. Statistical Science. 2002; 17:235–55.
6. Gupta K. New Delhi: Prentice Hall of India: Introduction to Data Mining with case studies. 2006.
7. Phua C, Alahakoon D, Lee V. Minority report in fraud detection: classification of skewed data. Sigkdd Explorations. 2004; 6:50-9.
8. Izadparast M, Farahi A, Fatahnezhad F, Tymvorpor B. The use of data mining techniques to predict the level of damage motor insurance customers, Quarterly Research Institute of Science and Technology of Information. 2012; 27:699–722.
9. Han J, Kamber M, Pei J. University of Illinois at Urbana-Champaign & Simon Fraser University: Data Mining, Concepts and Techniques, 3rd ed. 2010.
10. Saniee M. Tehran-Iran: Niyaz Danesh Publishing: Applied data Mining, first printing. 2012.
11. Karamizadeh F, Zolfagharifar A. Using the Clustering Algorithms and Rule-based of Data Mining to Identify Affecting Factors in the Profit and Loss of Third Party Insurance, Insurance Company Auto. Indian Journal of science and Technology. 2016 Feb; 9(7):145–52.
12. Delafrooz N, Farzanfar E. Determining the Customer Lifetime Value based on the Benefit Clustering in the Insurance Industry. Indian Journal of science and Technology. 2016 Jan; 9(1):1–8.