Kybernetes

Solving customer insurance coverage sales plan problem using a multi-stage data mining approach

Farshid Abdi, Kaveh Khalili-Damghani, Shaghayegh Abolmakarem,

Solving customer insurance coverage sales plan problem using a multi-stage data mining approach

Farshid Abdi, Kaveh Khalili-Damghani and Shaghayegh Abolmakarem
Department of Industrial Engineering, South Tehran Branch, Islamic Azad University, Tehran, Iran

Abstract

Purpose – Customer insurance coverage sales plan problem, in which the loyal customers are recognized and offered some special plans, is an essential problem facing insurance companies. On the other hand, the loyal customers who have enough potential to renew their insurance contracts at the end of the contract term should be persuaded to repurchase or renew their contracts. The aim of this paper is to propose a three-stage data-mining approach to recognize high-potential loyal insurance customers and to predict/plan special insurance coverage sales.

Design/methodology/approach – The first stage addresses data cleansing. In the second stage, several filter and wrapper methods are implemented to select proper features. In the third stage, K-nearest neighbor algorithm is used to cluster the customers. The approach aims to select a compact feature subset with the maximal prediction capability. The proposed approach can detect the customers who are more likely to buy a specific insurance coverage at the end of a contract term.

Findings – The proposed approach has been applied in a real case study of insurance company in Iran. On the basis of the findings, the proposed approach is capable of recognizing the customer clusters and planning a suitable insurance coverage sales plans for loyal customers with proper accuracy level. Therefore, the proposed approach can be useful for the insurance company which helps them to identify their potential clients. Consequently, insurance managers can consider appropriate marketing tactics and appropriate resource allocation of the insurance company to their high-potential loyal customers and prevent switching them to competitors.

Originality/value – Despite the importance of recognizing high-potential loyal insurance customers, little study has been done in this area. In this paper, data-mining techniques were developed for the prediction of special insurance coverage sales on the basis of customers' characteristics. The method allows the insurance company to prioritize their customers and focus their attention on high-potential loyal customers. Using the outputs of the proposed approach, the insurance companies can offer the most productive/economic insurance coverage contracts to their customers. The approach proposed by this study be customized and may be used in other service companies.

Keywords Data mining, Knowledge management, Feature selection algorithm, K-nearest neighbor algorithm, Insurance coverage sales

Paper type Research paper

1. Introduction

Market changes are collectively considered one of the main challenges facing insurance companies. Customers and competitors, two crucial factors for the insurance companies, are affected by these changes too. Due to the increasing number of insurance companies, and the

Kybernetes
© Emerald Publishing Limited
0368-492X
DOI 10.1108/K-07-2017-0244

variety of services provided by them, customers have the power to choose. Whenever the number of competitor increases, each company has some programs and new ways to attract more customers. The survival of insurance companies, perhaps most of all, depends on the sales. Insurance is an intangible product. The more insurance contract terms are annual, the greater of number of customers are lost at end of the year; therefore, identifying customers who are more likely to repurchase or renew their contracts is important forth insurance companies. Insurance marketers and vendors deal directly with the customers and sales network that is responsible for providing insurance services for individuals so that they will play a prominent role in identifying and attracting customers and making profit for the company.

Customers are valuable for the insurance companies, as service organizations and insurance companies rely on the relationship they share with them. Customer relationship management (CRM) system can be useful for companies to obtain information about their customers. Therefore, CRM system can help companies to identify valuable and profitable customers and to develop their relationship with them (Cheng and Chen, 2009). A comprehensive CRM system addresses customer identification, attraction and retention and relationship development with customers. The process of customer identification also includes the identification of the people who can be future customers or can produce the highest profits (Ngai *et al.*, 2009). Insurance companies collect data related to their customers. By using data-mining techniques, companies can extract useful knowledge from these data and gain an understanding of their customers and their needs by using this knowledge.

Previous studies applied various data-mining techniques in different insurance domains such as insurance premium pricing and determining the premium rate (Yeo *et al.*, 2002), product suggestion (Kumar and Singh, 2011), product development (Liao *et al.*, 2009), forecasting the probability and amount of loss (Lin, 2009), determining the total claim amounts in insurance (Dalkilic *et al.*, 2009), predicting life insurance installments (Shuang and We, 2011), detecting fraud (Tao *et al.*, 2012), customer segmentation (Thakur and Sing, 2013), targeting customers (Devale and Kulkarni, 2012), customer retention (Guelman *et al.*, 2012) and predicting customer preference (Balaji and Srivatsa, 2012).

Despite the importance of recognizing high-potential loyal insurance customers and insurance coverage sales plan, little study has been done in this area. In this paper, data-mining techniques were used to develop a model for the prediction of special insurance coverage sales on the basis of customers' characteristics. The prediction method presented in this article allows the insurance company to prioritize their customers and focus their attention on high-potential loyal customers and thereby increased profitability for the company.

The proposed approach can be remarkable from several aspects. On the one hand, due to the intense competition among insurance companies and to survive in this competitive market, managers are trying to attract more customers and create a long-term relationship with them. Competitors are ready to provide the same services and products with higher quality and lower prices and attract our lost customers. The previous studies showed that retaining the existing customers of the organization is much cheaper than new customer acquisition. A little increase in customer retention rate can lead to a remarkable increase in profits. Hence, paying attention to the retention of the customers is considered as one of the most crucial strategic components of profitability in the companies (Verbeke *et al.*, 2011).

Therefore, the proposed approach can be useful for the insurance company which helps them to identify their potential clients. Consequently, insurance managers can consider appropriate marketing tactics and appropriate resource allocation of the insurance company to their high-potential loyal customers and prevent switching them to competitors. On the other hand, not all customers have the same needs. Insurance companies should differentiate between their customers. In some cases, a particular insurance coverage is not suitable for all customers. Identifying the "right" prospects have a significant role in gaining profit. Based on the above description, it can be said that this approach can lead to changes in the market.

The knowledge extracted from CRM data is given to sales teams. Insurance vendors will be able to detect target customers leading to an increase in sales by making use of the extracted information (obtaining from data mining). In this article, data-mining techniques are used to improve CRM and identify valuable customers. A model is proposed to predict whether a specific insurance coverage will be bought by a customer or not. This prediction is based on customers' characteristics. In fact, this model enables insurance companies to identify the people who are more likely to buy insurance coverage. The model of this article can be a very effective and helpful tool for companies, because of the importance of sales for insurance companies. This paper is organized as follows: in Section 2, the literature of past works is presented. In Section 3, the theoretical background is presented. In Section 4, research methodology of the paper is explained. In Section 5, experimental approach is presented. In Section 6, results are discussed, and in Section 7, conclusions are represented.

2. Literature of past works

The aim of feature selection (FS) algorithms is to exclude irrelevant or redundant features to increase precision in classification and clustering (Tsai *et al.*, 2013). FS methods are divided into four broad groups: filter methods, wrapper methods, embedded methods and hybrid methods (Moradkhani *et al.*, 2015; Guyon and Elisseeff, 2003). The use of these techniques has been previously investigated by several works. In a study, authors examined several FS techniques and proposed a hybrid system with genetic algorithm (GA) and artificial neural network (NN) for FS at retail credit risk data set (Oreski *et al.*, 2012). Oreski and Oreski (2014) also used the hybrid genetic algorithm with NNs in another study to identify an optimum feature subset and increased the accuracy of the classification. In a study, several filter methods (relief, correlation-based FS, fast correlated-based filter and INTERACT) were applied to the artificial data sets (Sánchez-Maroño *et al.*, 2007).

An improved classical Gini Index algorithm was suggested by the authors of a study to improve the performance of text categorization (Shang *et al.*, 2007). In several studies, authors used wrapper methods, i.e. sequential backward selection, sequential forward selection and optimized selection (evolutionary) for FS (Panthong and Srivihok, 2015; Maldonado and Weber, 2009). A constructive approach to FS based on wrapper methods and sequential search strategy was presented in a study for FS (Kabir *et al.*, 2010). This approach used a feed forward NN as a training model. Several studies have used hybrid techniques. Zhao *et al.* (2015) proposed a two-stage method combining information gain and binary particle swarm optimization to find the optimal feature set. The combination of instance-based learning to generate candidate feature subsets (CFSs) and a cooperative subset search algorithm to evaluate the CFSs was used in a study to select a subset of features (Brahim and Limam, 2016). In a study, the author proposed a two-stage FS method including information gain, GA and principal component to increase the degree of accuracy of classification (Uguz, 2011). Hsu *et al.* (2011) introduced a hybrid FS method by combining filters and wrappers.

Two hybrid FS techniques, called BDE-X Rank and BDE-X Rank (BDE – binary differential evolutionary), consisting of two stages were proposed. Both algorithms combined a binary differential evolutionary algorithm with a ranked-based filter method (Apolloni *et al.*, 2016). In another study, a two-stage FS method was proposed again. In the filter phase, a symmetrical uncertainty criterion was used to create weighted features, and in the wrapper phase, fuzzy imperialist competition algorithm and incremental wrapper subset selection with replacement were used to search efficient feature subset (Moradkhani *et al.*, 2015). Therefore, as mentioned in the previous articles, the general purpose of feature subset selection is to achieve a mechanism to remove some irrelevant and redundant features and extract optimal number of features to increase efficiency and accuracy of classification. In the present study, the FS method was used to select the influential features to achieve the model with compact subset of customers' features and acceptable accuracy.

3. Theoretical background

3.1 Data preparation

Here, some data pre-processing and cleansing methods used in the present study are briefly described.

3.1.1 Local outlier factor. Computed for each record in a data set to represent the outlier degree of a record. The local outlier factor (LOF) is a method that uses the nearest neighborhood concept to identify outliers (Bai *et al.*, 2016; Ye *et al.*, 2016).

3.1.2 Distance-based method. In this method, the top n records which are the farthest from the K-nearest neighbor are considered as outliers (Ghoting *et al.*, 2008; Ramaswamy *et al.*, 2000).

3.1.3 Missing value imputation. In this method, the value of missing data is estimated by using a suitably supervised learning algorithm, i.e. K-nearest neighbor algorithm (Sim *et al.*, 2016).

3.2 Feature selection algorithms

3.2.1 Filter methods.

3.2.1.1 Information gain. This index is used as the concept of information gain to determine the correlation between each feature and class variable (Oreski *et al.*, 2012).

3.2.1.2 Gini Index. Obtaining "Gini Index" is another filter method. This index is used to measure the ability of a feature to identify the classes. Using Gini Index, the noises of training records in D are calculated as (Zhao *et al.*, 2015):

$$Gin(D) = 1 – \sum_{i=1}^{m} p_i^2 \tag{1}$$

Where p_i is the probability that a tuple (in D) belongs to class C_i.

Gini Index of each feature is calculated independently, and the top k features with the smallest Gini Index are selected.

3.2.1.3 Correlation. In this method, features are weighted on the basis of correlations between features. Those features having the least degree of correlation take the maximum weight. Correlation between Features A and B is calculated as (Jiang and Wang, 2016):

$$\rho(A,B) = \frac{Cov(A,B)}{\sqrt{Var(A)Var(B)}} \tag{2}$$

3.2.2 Wrapper methods.

3.2.2.1 Genetic algorithm. A GA is an optimization method that uses the idea of evolutionary theory, and it is capable to search efficiently in large spaces (Welikala *et al.*, 2015; Uguz, 2011).

3.2.2.2 Forward. This algorithm starts with a null feature set, and for every step, the most effective feature satisfying some evaluation metrics are added to the present features (Oreski *et al.*, 2012; Panthong and Srivihok, 2015).

3.3 Classification algorithm

3.3.1 K-nearest neighbor. K-nearest neighbor classification algorithm is done on the basis of similarity, by comparing similar testing records and training records. All training records are kept in an n-dimensional space, and K-nearest neighbor finds a record which is as similar as possible to an unknown record. The similarity of records can usually be defined by Euclidean distance (Han and Kamber, 2006; Larose and Larose, 2014).

4. Methodology

4.1 Data set description

We have used the Cochran formula to test the statistical significance of the sample. Based on unlimited statistical population and at the significant level equal to 0.05, the sufficient sampling size is equal to 384. The data set used in this study contains the information of 478 customers of an insurance company in Iran[1]. A sample with 478 members easily fulfill the requirement of Cochran sampling formula. In all, 70 customers buy a particular insurance coverage (Buyer) and 408 customers do not buy this type of insurance coverage (non-buyer). The variables and the types of them used in the proposed model and are represented in Table I.

4.2 The proposed model

The proposed model, as shown in Figure 1, is a three-stage approach:

(1) data cleansing;

(2) data pre-processing; and

(3) modeling.

	Variables related to customers' characteristics	
Row	Variables	Type
1	Age	Nominal
2	Family members	Numerical
3	Education	Nominal
4	Job	Nominal
5	Home type	Numerical
6	Units	Numerical
7	Year	Numerical
8	Owner	Binominal
9	Income	Nominal
10	Region	Numerical
11	Cigarette	Binominal
12	Have child?	Binominal
13	Read newspaper?	Nominal
14	Using internet?	Nominal
15	Value of home	Nominal
16	Point of buy	Nominal

Table I. Variables related to customers' characteristics and their types

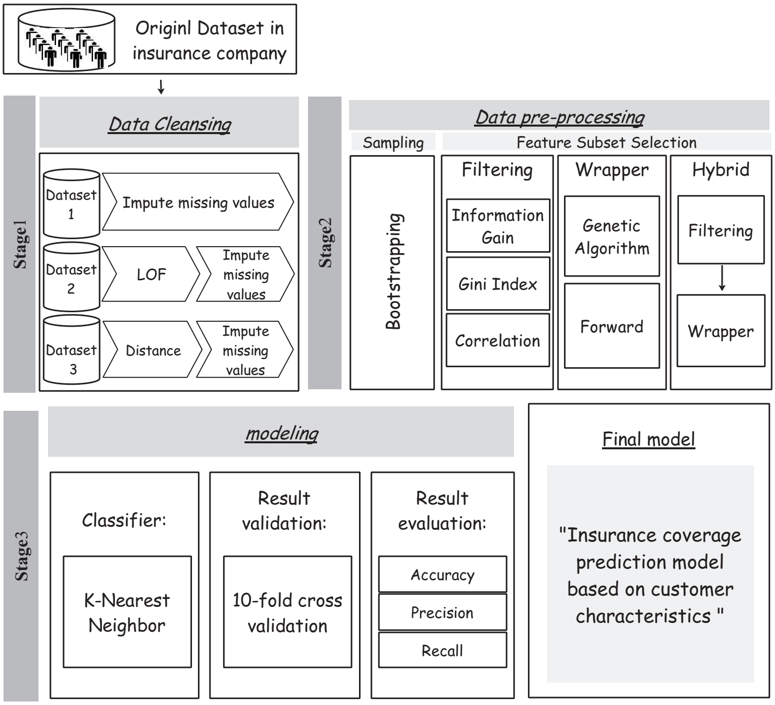


Figure 1.
Three-stage methodology of research

In the first stage, data are cleansed. In the second stage, after sampling, several filter and wrapper techniques, including information gain, Gini Index, correlation, forward selection and GA, are used separately and in hybrid form. In the third stage, the selected features are assumed as the input of K-nearest neighbor classification algorithm. Finally, the accuracy of the model is compared and the best model is selected.

5. Experimental approach

All algorithms of this research are implemented using Rapid Miner software, version 5.3, running on a 2.4 GHz CPU with 4 GB RAM and windows 7-32-bit operating system.

5.1 First stage: data cleansing

In the first stage, the data are cleansed so as to detect and remove the outliers by using the distance-based method and computing LOF for each record, and missing values are imputed using a K-nearest neighbor algorithm. As represented in Figure 1, three different data sets are obtained from data cleansing, namely, Data set 1, Data set 2 and Data set 3.

5.2 Second stage: bootstrapping and feature subset selection

In this stage, there is an imbalance in the number of records in classes; therefore, bootstrapping method is used to solve this issue. Because the main objective of this study is to present a model to predict the specific insurance coverage sales on the basis of customers’ characteristics, FS techniques are used to find the most proper features. In the present study, wrapper techniques (GA and forward method) and filter techniques (Gini Index, correlation and information gain) are used separately and in hybrid form.

5.2.1 Feature selection by filter methods. Information gain, Gini Index and correlation as the FS techniques provide a weight for each feature. The filter approach assigns weights to each feature independent of the classification algorithm (Maldonado and Weber, 2009).

Figure 2 shows the overall process of FS done by filter methods in the present study. As shown in Figure 2, based on the weights given to each feature by the algorithm, K features with the highest weights are put into the model. In this study, K is investigated for the values of 3, 4, 5 and 6. In each case, the accuracy, precision and recall of the model are measured.

5.2.2 Feature selection by wrapper methods. In addition to filter FS method, there are some methods in which evaluation function is based on classifier error rate measures. These are called wrapper methods. In these methods, a classification function is used to evaluate the usefulness of subsets of features. Figure 3 shows the process of FS in wrapper methods. As shown in Figure 3 of this study, K-nearest neighbors are used as fitness function in GA and forward selection techniques. K investigated for the values of 2, 5 and 10.

5.2.3 Hybrid feature selection. In present study, to examine whether the combination of the filter and wrapper methods increases the accuracy of the classification, the hybrid form of these methods is issued too.

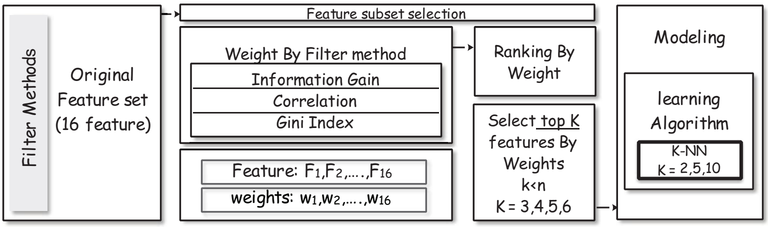
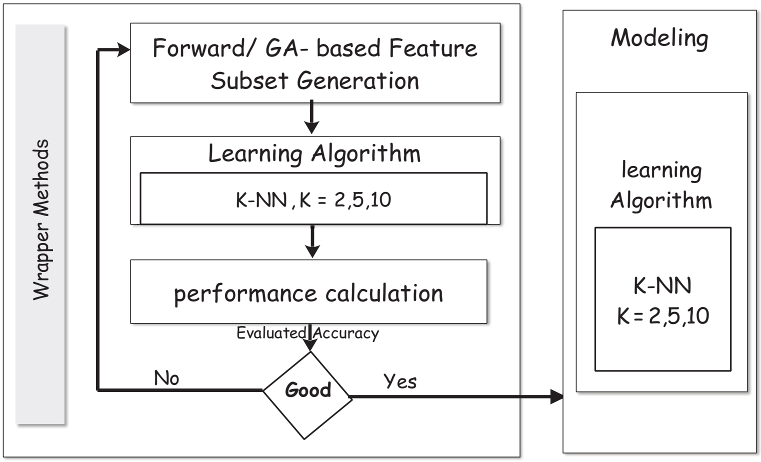


Figure 2. Overall process of feature selection by filter methods in this study



Sources: Oreski *et al*. (2012); Huang *et al*. (2007)

Figure 3. Overall process of feature selection in wrapper methods

First, filter methods (Gini Index, correlation and information gain) are used for weighting features. In the second stage, ten features which are given the highest weights by the filter method are entered into the forward method. In forward method, K-nearest neighbor algorithm is used as the fitness function. Finally, features selected by forward method are assumed to be an input of K-nearest neighbor algorithm in the modeling stage. Figure 4 shows the overall design of hybrid FS method in this study.

5.3 Modeling

The efficiency of the selection of variables was examined with accuracy, precision and recall of the class prediction. K-nearest neighbor is used as a classifier. A 10-fold cross validation is applied to evaluate the performance of the data-mining models.

5.4 Evaluation of results

In this section, commonly used metrics in machine learning for evaluation such as accuracy, precision and recall are selected to evaluate the classification performance. "Confusion Matrix" is used to achieve this objective. Table II shows the confusion matrix when data are presented in two classes (Wang and Ma, 2012). Recall shows the portion of the true positives which are correctly diagnosed as positive, and precision demonstrates the portion of the true positives diagnosed by the classifier, which are really true positive (Harris, 2013).

The classification accuracy, precision and recall are depicted, respectively, in equations (3), (4) and (5):

$$Accuracy = \frac{TN + TP}{TP + TN + FP + FN} \tag{3}$$

$$Precision = \frac{TN}{TN + FN} \tag{4}$$

$$\mathrm{Re}call = \frac{TN}{TN + FN} \tag{5}$$

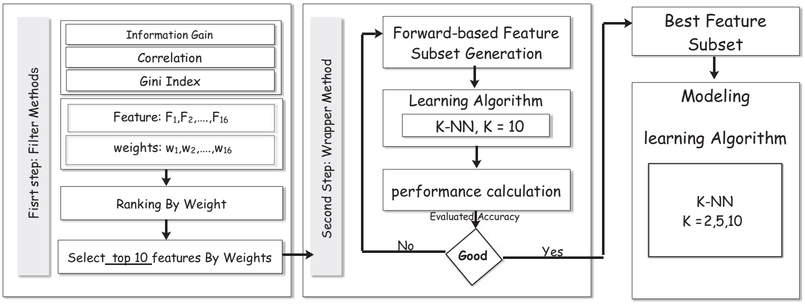


Figure 4. Overall design of hybrid feature selection method in the present study

Table II. Confusion matrix

		Actual	
		Positive	Negative
Prediction	Positive	True negative (TP)	False positive (FP)
	Negative	False negative (FN)	True negative (TN)

6. Results and discussion

Table III presents the results of modeling by K-nearest neighbor classifier in three different modes of data cleansing (mentioned in Section 5.1) and FS based on filter and wrapper methods. According to Table III, Model 2 has higher accuracy in the case of GA FS method. The values of accuracy, precision and recall are 95.58, 92.67 and 74.33 per cent, respectively. As we see, the number of the selected features in this model is 8.

Model 1 achieved the next rank, whose accuracy, precision and recall are 94.88, 91.67 and 70.33 per cent, respectively. In this case, three features are selected by forward selection method. As presented in Table III, in the second mode of data cleansing (Data set 2), Model 7 in the second mode of data cleansing has better performance than the other models in this mode. In Model 7, FS performs based on GA, and the number of selected features is 5. Model 11, by using forward selection method and selecting five features, achieved 93.12 per cent accuracy and performed better than other models in the third mode of data cleansing. Model 9 is the model giving the worst classification performance, considering the lowest accuracy, precision and recall measures. In this model, Gini Index method is used for FS. With respect to Table III, in all modes of data cleansing, wrapper methods gain higher accuracy than filter methods.

As mentioned before, the model in this paper aims to select a compact feature subset with the maximal predictive capability. So, it could be stated that Model 1 (by selecting three features and 94.88 per cent accuracy) performs better than Model 2 (by selecting eight features and 95.58 per cent accuracy). So, Model 1 is the selected model of this research.

Figure 5 shows the comparison of classification results based on the FS technique and the number of selected features for three data sets generated from data cleansing. The forward method performs better in Data sets 1 and 3.

Table IV shows the features selected in each modeling. As displayed in Table IV, the selected features in Model 1 are “Family member”, “Units” and “Region”. As can be seen, the family member is selected by 5 models, units by 8 models and region by 12 models.

Figure 6 draws a comparison between accuracy and the number of selected features of the top five models of Table III, based on their accuracies.

	No. of		No. of		K-NN, K = 2	
Data set	model	FSS techniques	features	Accuracy (%)	Precision (%)	Recall (%)
1	1	Forward, K-NN, K = 2	3	94.88 [2.03]	91.67 [10.54]	70.33 [18.04]
	2	GA, K-NN, K = 2	8	95.58 [3.02]	92.67 [11.72]	74.33 [16.80]
	3	Gini Index	4	93.95 [2.37]	83.14 [11.89]	72.00 [16.81]
	4	Information gain	5	91.40 [3.13]	67.93 [10.15]	67.33 [17.31]
	5	Correlation	5	91.40 [3.13]	67.93 [10.15]	67.33 [17.31]
2	6	Forward, K-NN, K = 2	7	92.75 [4.14]	77.50 [30.27]	61.00 [9.40]
	7	GA, K-NN, K = 2	5	93.22 [3.87]	85.83 [20.77]	57.67 [21.55]
	8	Gini Index	6	90.40 [3.59]	62.80 [24.49]	57.33 [27.72]
	9	Information gain	6	91.11 [3.75]	68.33 [16.89]	62.67 [22.40]
	10	Correlation	6	90.40 [3.59]	62.80 [24.49]	57.33 [27.72]
3	11	Forward, K-NN, K = 2	5	93.12 [1.95]	0.83 [14.17]	57.29 [12.06]
	12	GA, K-NN, K = 2	8	93.12 [3.26]	90.17 [12.53]	56.95 [22.61]
	13	Gini Index	6	91.22 [2.80]	69.17 [10.52]	67.71 [12.61]
	14	Information gain	6	91.69 [3.23]	72.71 [13.89]	67.71 [12.61]
	15	Correlation	6	87.41 [3.00]	59.29 [19.42]	43.52 [17.47]

Table III. Results of modeling based on different modes of data cleansing and single feature selection techniques

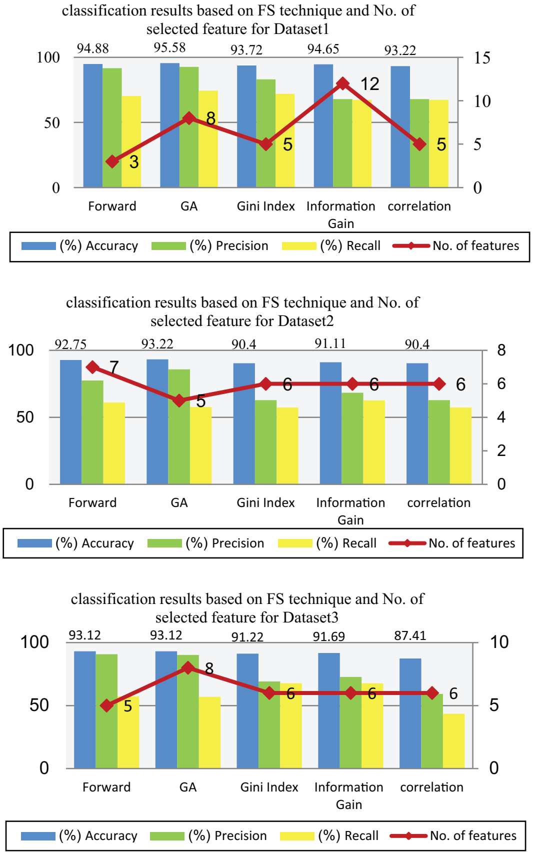


Figure 5.
Comparison of the results based on FS technique and number of the selected features

Table V represents the performance of Model 1 in the form of the confusion matrix. As shown in the table, among 58 people who are the actual buyers, 41 people are predicted as buyers. Also, 89.13 per cent of the actual buyers are correctly categorized.

In the best model of study (Model 1), modeling is done with other classifiers, in addition to classification with K-NN algorithm, i.e. classifiers such as decision tree, Naïve Bayes, bagging K-NN, bagging DT and bagging NB. By comparing the results, it can be shown that the K-NN algorithm has better performance than other algorithms. The results are represented in Table VI. As can be seen in Table VI, the best accuracy (94.88 per cent) is achieved in the case of modeling by K-NN and bagging K-NN.

Figure 7 also shows the receiver operating characteristic (ROC) curve for the mentioned classifiers. Vertical axes show the true positive rate, and as a result, the horizontal axes show the false positive rate. The accuracy is measured by the area under the ROC curve. An area of 1 represents the best classifier, and an area of 0.5 represents the classifier with the worst performance. So according to Figure 7, K-NN performs better and NB and bagging NB are the two methods with the worst classification performance.

In this paper, the filter methods, wrapper methods and hybrid form of them are used. Features are selected in two stages. In the first stage, features of customers are weighted by

Table IV. Features selected in each modeling

Model	1 Age	2 Family members	3 Education	4 Job	5 Home type	6 Units	7 Year	8 Owner	9 Income	10 Region	11 Cigarette	12 Have children?	13 Read newspaper?	14 Using internet?	15 Value of home	16 Point of buy	No. of selected features
1		×				×				×							3
2	×	×	×		×				×			×			×	×	8
3		×	×			×				×							4
4						×				×		×	×			×	5
5						×				×		×	×			×	5
6			×			×				×				×		×	5
7				×			×	×	×				×				5
8	×	×					×		×	×							6
9	×					×			×	×			×		×		6
10	×	×					×		×	×						×	6
11			×							×		×			×	×	5
12	×		×	×						×		×	×		×	×	8
13	×		×			×	×			×					×		6
14	×					×	×	×		×					×		6
15					×		×	×					×	×	×		6
Number of models select the feature	7	5	6	2	2	8	6	3	5	12	0	5	6	2	7	7	

filter methods (Gini Index, correlation and information gain). In the second stage, forward method is used to select features. In the forward method, K-NN classification algorithm is used as the fitness function. The features selected in the second stage are transferred into the modeling stage. In this stage, K-NN algorithm is used as classifications algorithm. To evaluate the models, the "Accuracy", "Recall" and "Precision" of the models are calculated. The results are presented in Table VII. As can be seen, the hybrid form does not improve the results. Figure 8 draws a comparison between the numbers of the selected features, accuracy, precision and recalls of the hybrid FS techniques.

The data sharing is a proven concept and can have positive effects including enhancing cooperation with customers, obtaining new partners with the desired reputation and qualification.

Also, the authorization of sharing the positive experiences between business partners can play an important role in the business partner reputation. Of course, considering the legal and ethical aspects of data sharing is also important, data sharing should be done with legal principles.

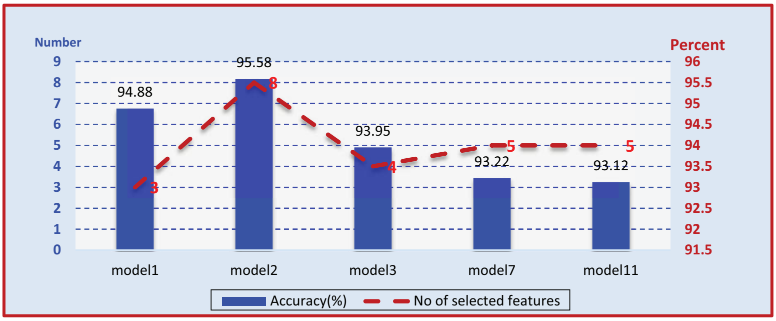


Figure 6.
Comparison of accuracy and number of selected features of top five models

Table V.
Confusion matrix for model 1

Test results	Actual Non-buyers	Buyers	Class precision (%)
Pred. 0	367	17	95.57
Pred. 1	5	41	89.13
Class recall (%)	98.66	70.69	

Table VI.
Comparison the performance of the best model of study and several classifiers

Classifiers	Accuracy (%)	Precision (%)	Recall (%)
KNN	94.88 [2.03]	91.67 [10.54]	70.33 [18.04]
DT	90.93 [2.19]	66.48 [10.19]	68.67 [14.00]
NB	85.81 [1.63]	20.00	1.67 [5.00]
Bagging KNN	94.88 [2.03]	91.67 [10.54]	70.33 [18.04]
Bagging DT	90.93 [2.19]	66.48 [10.19]	68.67 [14.00]
Bagging NB	85.81 [1.63]	20.00	1.67 [5.00]

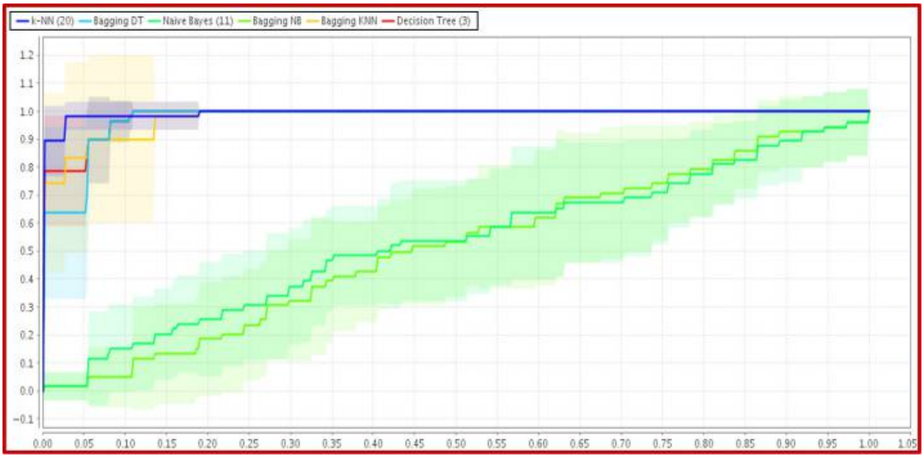


Figure 7.
ROC Curve

Data sharing between insurance companies can be useful in some aspects as follows. In some cases, insurance companies need to know the records of clients in other insurance companies. The possibility for insurance company to access the required records from the client can prevent many fraud cases in the insurance industry. Identifying and meeting customer needs, informing customers about products and opportunities, increasing access to various insurance services and coverage and increasing customer satisfaction can be considered as other benefits of sharing data and information between insurance companies.

7. Conclusion

One of the main issues in the managing of the insurance companies is CRM. The use of information and communication technology and data-mining tools to achieve optimal CRM led to an ongoing relationship with their customers and ultimately to customers' satisfaction and loyalty. Loyal customers had a considerable profit for insurance companies. Insurance contracts were often issued for a particular period of time (usually a year). The end of the contract term can be considered as the end of the customer's life. So the identification of customers who were likely to repurchase and renew their insurance contract was important for insurance companies. Retaining existing customers and identifying future customers can be considered as a competitive advantage for insurance companies.

In the present study, data-mining techniques were used to make a model for the prediction of special insurance coverage sales on the basis of customers' characteristics. The filter and wrapper methods were used for feature subset selection, both separately and in the hybrid form. By using these techniques, a subset of influential features was selected and put into the learning algorithm. In this paper, the forward method had better results. In modeling stage, K-nearest neighbor algorithm was used as the classification algorithm. The best model was the one which could produce the highest degree of accuracy with the least number of features.

Insurance companies can identify the people who were more likely to become their customers by using the model proposed in this study. Identifying these people helped the insurance companies to do more effective marketing activities and more purposeful advertisement to retain existing customers and attract new customers. Also, it allowed the

Table VII. The best results of the modeling in hybrid feature selection technique

		Feature subset selection					K-NN, K = 2	
Data set	No. of hybrid model	Step 1	No. of feature	Step 2	No. of feature	Accuracy (%)	Precision (%)	Recall (%)
1	1	Gini Index	10	Forward	3	94.88 [2.03]	91.67 [10.54]	70.33 [18.04]
	2	Correlation	10	Forward	7	93.02 [3.29]	79.12 [16.36]	70.67 [13.15]
	3	Information Gain	10	Forward	3	94.88 [2.03]	91.67 [10.54]	70.33 [18.04]
2	4	Gini Index	10	Forward	5	92.04 [4.10]	79.17 [23.35]	56.00 [22.70]
	5	Correlation	10	Forward	4	90.65 [5.01]	67.08 [31.21]	52.33 [27.08]
	6	Information Gain	10	Forward	6	89.94 [3.02]	63.39 [13.69]	63.00 [19.63]
3	7	Gini Index	10	Forward	4	91.45 [3.22]	82.33 [18.98]	50.62 [20.34]
	8	Information Gain	10	Forward	4	91.45 [3.22]	82.33 [18.98]	50.62 [20.34]

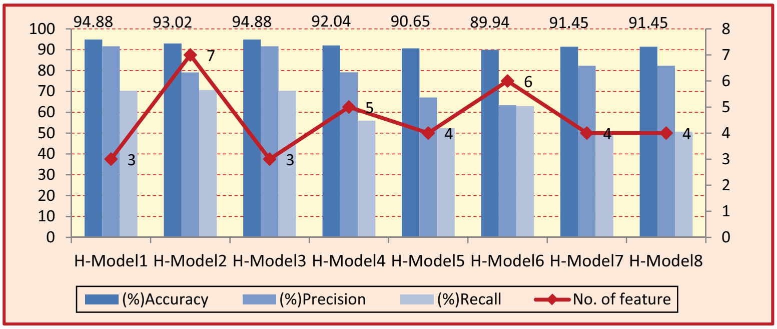


Figure 8.
Comparison of the models

marketing and sales teams to devote the time and resource of the company to valuable customers who were more likely to purchase insurance coverage. In this way, it produced more profit for company and reduced marketing costs.

In the challenging business market, perceiving the customers' behavior can be an important critical success factor. This can help companies to plan new strategies while competing in a dynamic situation. The proposed approach of this study can be customized and applied in a wide range of problems regarding customers in service organizations such as banks, hospitals and libraries.

Although, one of the main limitations of this study was its bounding limitations of certainty of future situations and on the basis of iterative behavior of the customers in the future. On the other hand, if the behavior of the customers is repeated in the future, the proposed approach of this study incorporating the historical data can help to identify the most suitable insurance coverage sales plan for each customer. If some uncertainty is occurred in the future, the behavior of the customers may be differing from the patterns achieved form historical data. So, development of the proposed approach of this study for uncertain situations modeled through fuzzy sets, random variables or robust modeling may be an interesting future research direction.

Note

1. The name of company is not presented here for sake of anonymity.

References

Apolloni, J., Leguizamon, G. and Alba, E. (2016), "Two hybrid wrapper-filter feature selection algorithms applied to high dimensional microarray experiments", *Applied Soft Computing*, Vol. 38 No. 1, pp. 922-932.

Bai, M., Wang, X., Xin, J. and Wang, G. (2016), "An efficient algorithm for distributed density-based outlier detection on big data", *Neuro Computing*, Vol. 181 No. 1, pp. 19-28.

Balaji, S. and Srivatsa, S.K. (2012), "Data mining model for insurance trade in CRM system", *International Journal of Advanced Research in Computer Science and Software Engineering*, Vol. 2 No. 4, pp. 232-235.

Brahim, A.B. and Limam, M. (2016), "A hybrid feature selection method based on instance learning and cooperative subset search", *Pattern Recognition Letters*, Vol. 69 No. 1, pp. 28-34.

Cheng, C.-H. and Chen, Y.-S. (2009), "Classifying the segmentation of customer value via RFM model and RS theory", *Expert Systems with Applications*, Vol. 36 No. 3, pp. 4176-4184.

Dalkilic, T.E., Tank, F. and Sanli Kula, K. (2009), "Neural networks approach for determining total claim amounts in insurance", *Insurance: Mathematics and Economics*, Vol. 45 No. 2, pp. 236-241.

Devale, A.B. and Kulkarni, R.V. (2012), "Applications of data mining techniques in life insurance", *International Journal of Data Mining & Knowledge Management Process*, Vol. 2 No. 4, pp. 31-40.

Ghoting, A., Parthasarathy, S. and Otey, M.E. (2008), "Fast mining of distance-based outliers in high-dimensional datasets", *Data Mining and Knowledge Discovery*, Vol. 16 No. 3, pp. 349-364.

Guelman, L., Guillén, M. and Pérez-Marín, A.M. (2012), *Random Forests for Uplift Modeling: An Insurance Customer Retention Case*, Springer-Verlag Berlin Heidelberg, pp. 123-133.

Guyon, I. and Elisseeff, A. (2003), "An introduction to variable and feature selection", *Journal of Machine Learning Research*, Vol. 3 No. 1, pp. 1157-1182.

Han, J. and Kamber, M. (2006), *Data Mining: Concepts and Techniques*, 2nd ed., Morgan Kaufmann Publishers.

Harris, T. (2013), "Quantitative credit risk assessment using support vector machines: broad versus narrow default definitions", *Expert Systems with Applications*, Vol. 40 No. 1, pp. 4404-4413.

Hsu, H.-H., Hsieh, C.-W. and Lu, M.-D. (2011), "Hybrid feature selection by combining filters and wrappers", *Expert Systems with Applications*, Vol. 38 No. 1, pp. 8144-8150.

Huang, C.-L., Chen, M.-C. and Wang, C.-J. (2007), "Credit scoring with a data mining approach based on support vector machines", *Expert Systems with Applications*, Vol. 33 No. 4, pp. 847-856.

Jiang, S-Y. and Wang, L-X. (2016), "Efficient feature selection based on correlation measure between continuous and discrete features", *Information Processing Letters*, Vol. 116 No. 2, pp. 203-215.

Kabir, M.M., Islam, M.M. and Murase, K. (2010), "A new wrapper feature selection approach using neural network", *Neurocomputing*, Vol. 73 Nos 16/18, pp. 3273-3283.

Kumar, P. and Singh, D. (2011), "Integrating data mining and AHP for life insurance product recommendation", *Computational Intelligence and Information Technology*, Springer-Verlag Berlin Heidelberg, pp. 596-602.

Larose, D.T. and Larose, C.D. (2014), *Discovering Knowledge in Data: An Introduction to Data Mining*, 2nd ed., John Wiley & Sons, Hoboken, NJ.

Liao, S.-H., Chen, Y.-N. and Tseng, Y.-Y. (2009), "Mining demand chain knowledge of life insurance market for new product development", *Expert Systems with Applications*, Vol. 36 No. 1, pp. 9422-9437.

Lin, C. (2009), "Using neural networks as a support tool in the decision making for insurance industry", *Expert Systems with Applications*, Vol. 36 No. 3, pp. 6914-6917.

Maldonado, S. and Weber, R. (2009), "A wrapper method for feature selection using support vector machines", *Information Sciences*, Vol. 179 No. 13, pp. 2208-2217.

Moradkhani, M., Amiri, A., Javaherian, M. and Safari, H. (2015), "A hybrid algorithm for feature subset selection in high-dimensional datasets using FICA and IWSSr algorithm", *Applied Soft Computing*, Vol. 35 No. 1, pp. 123-135.

Ngai, E.W.T., Xiu, L. and Chau, D.C.K. (2009), "Application of data mining techniques in customer relationship management: a literature review and classification", *Expert Systems with Applications*, Vol. 36 No. 1, pp. 2592-2602.

Oreski, S. and Oreski, G. (2014), "Genetic algorithm-based heuristic for feature selection in credit risk assessment", *Expert Systems with Applications*, Vol. 41 No. 4, pp. 2052-2064.

Oreski, S., Oreski, D. and Oreski, G. (2012), "Hybrid system with genetic algorithm and artificial neural networks and its application to retail credit risk assessment", *Expert Systems with Applications*, Vol. 39 No. 16, pp. 12605-12617.

Panthong, R. and Srivihok, A. (2015), “Wrapper feature subset selection for dimension reduction based on ensemble learning algorithm”, *Procedia Computer Science*, Vol. 72 No. 1, pp. 162-169.

Ramaswamy, S., Rastogi, R. and Shim, K. (2000). “Efficient algorithms for mining outliers from large datasets”, *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 427-438.

Sánchez-Maroño, N., Alonso-Betanzos, A. and Tombilla-Sanromán, M. (2007). “Filter Methods for Feature Selection – A Comparative Study”, *Proceedings of the 8th International Conference on Intelligent Data Engineering and Automated Learning*, pp. 178-187.

Shang, W., Huang, H., Zhu, H., Lin, Y., Qu, Y. and Wang, Z. (2007), “A novel feature selection algorithm for text categorization”, *Expert Systems with Applications*, Vol. 33 No. 1, pp. 1-5.

Shuang, P. and We, L. (2011), *The Early Warning of Life Insurance Company Based on BP Artificial Neural Network*, Springer-Verlag Berlin Heidelberg, pp. 30-38.

Sim, J., Kwon, O. and Lee, K.C. (2016), “Adaptive pairing of classifier and imputation methods based on the characteristics of missing values in data sets”, *Expert Systems with Applications*, Vol. 46 No. 1, pp. 485-493.

Thakur, S.S. and Sing, J.K. (2013). “Mining customer’s data for vehicle insurance prediction system using k-Means clustering - An application”, *International Journal of Computer Applications in Engineering Sciences*, Vol. 3 No. 4, pp. 148-153.

Tsai, C.-F., Eberle, W. and Chu, C.-Y. (2013), “Genetic algorithms in feature and instance selection”, *Knowledge-Based Systems*, Vol. 39 No. 1, pp. 240-247.

Uguz, H. (2011), “A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm”, *Knowledge-Based Systems*, Vol. 24 No. 1, pp. 1024-1032.

Verbeke, W., Martens, D., Mues, C. and Baesens, B. (2011), “Building comprehensible customer churn prediction models with advanced rule induction techniques”, *Expert Systems with Applications*, Vol. 38 No. 1, pp. 2354-2364.

Wang, G. and Ma, J. (2012), “A hybrid ensemble approach for enterprise credit risk assessment based on support vector machine”, *Expert Systems with Applications*, Vol. 39 No. 5, pp. 5325-5331.

Welikala, R.A., Fraz, M.M., Dehmeshki, J., Hoppe, A., Tah, V., Mann, S., Williamson, T.H. and Barman, S.A. (2015), “Genetic algorithm based feature selection combined with dual classification for the automated detection of proliferative diabetic retinopathy”, *Computerized Medical Imaging and Graphics*, Vol. 43 No. 1, pp. 64-77.

Ye, Z., Cao, H., Zhang, Y. and Jia, L. (2016), “Outlier factor based partitional clustering analysis with constraints discovery and representative objects generation”, *Neurocomputing*, Vol. 173 No. 1, pp. 1538-1553.

Yeo, A.C., Smith, K.A., Willis, R.J. and Brooks, M. (2002), “A mathematical programming approach to optimise insurance premium pricing within a data mining framework”, *Journal of the Operational Research Society*, Vol. 53 No. 11, pp. 1197-1203.

Zhao, X., Li, D., Yang, B., Chen, H., Yang, X., Yu, C. and Liu, S. (2015), “A two-stage feature selection method with its application”, *Computers and Electrical Engineering*, Vol. 47 No. 1, pp. 114-125.

About the authors

Farshid Abdi is an Assistant Professor of Industrial Engineering at the School of Industrial Engineering, South-Tehran Branch, Islamic Azad University, Tehran, Iran. He received his PhD in Industrial Management in 2005. His research interests are data mining, customer relationship management, business intelligence, structural equation modeling, total quality management, service management and productivity management. He has published several scientific papers in international journals and conferences.

Kaveh Khalili–Damghani is an Associate Professor at the Islamic Azad University, South Tehran Branch, Faculty of Industrial Engineering, the Department of Socio-Economics Systems, Tehran, Iran. He received a PhD degree in the field of Industrial Engineering in 2009, and a PhD degree in the field of Industrial Management in 2012. His research interests include computational intelligence, applied operations research, soft computing methods and quantitative methods of decision science. He is an Associate Editor of six international scientific journals. He has published more than 90 international research papers. Kaveh Khalili-Damghani is the corresponding author and can be contacted at: kaveh.khalili@gmail.com

Shaghayegh Abolmakarem is a PhD Candidate of Industrial Engineering at the School of Industrial Engineering, South-Tehran Branch, Islamic Azad University, Tehran, Iran. Her research interests are computational intelligence, soft computing, data mining and service organization performance management.