Research Article

Construction of a Recommendation Method for Financial Insurance Products Based on Machine Learning

Bingyan Wu,1 Xiaoqing An,1 Wei Sun,2 Xue Chang,3 and Huadong Su4

1School of Economics, Jinan University, Guangzhou, Guangdong 510632, China
2School of Economics, Anyang Normal University, Anyang, Henan 455001, China
3University of Nottingham Ningbo China, Ningbo, Zhejiang 315000, China
4Department of Community & International Development, Yeungnam University, Gyeongsan-si, Gyeongsangbuk-do 38541, Republic of Korea

Correspondence should be addressed to Xiaoqing An; anxiaoq@stu2019.jnu.edu.cn

Received 5 March 2022; Accepted 5 April 2022; Published 30 April 2022

Academic Editor: Song Jiang

Copyright © 2022 Bingyan Wu et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

In the era of the rise of big data, if insurance companies can effectively and reasonably use existing data to tap more potential information, it can not only improve work efficiency but also achieve precise marketing to customers, thereby saving costs. At the same time, China pays more and more attention to financial insurance, so it is of great significance to use customer information to explore the purchase behavior of financial insurance. This paper mainly analyzes the factors that affect the purchase of financial insurance from the aspects of individuals and families and provides references for insurance marketing. This paper selects the results of the China Comprehensive Social Survey in 2020 as the empirical sample data of this paper and sets whether to purchase financial insurance as the target variable. The characteristic variables were screened, and the 10 variables selected after screening were finally used to build the model. Next, the data are divided into training set and test set, and a decision tree learning model is established on the two data sets at the same time. The classification results of the model are evaluated according to the classification evaluation indicators. Finally, the decision tree analysis results provide suggestions and strategies for the construction of insurance marketing recommendation methods.

1. Introduction

Financial insurance is a collective term for finance and insurance. Finance refers to the financing of monetary funds and is a general term for the adjustment activities of currency circulation centered on banks. Banks and other financial institutions operate deposits, loans, savings, bill discounting, foreign exchange, settlement, trust, investment, financial leasing, issuance of negotiable securities, and so on, all of which belong to the general laws of financial business development in accordance with the market economy [1, 2].

Now, we are in the era of “Internet +” and big data. Behind these data, there are many business opportunities and information. For enterprises, it is very necessary to find effective information to help enterprises develop from these large amounts of data [3, 4]. Only by mining the required relevant information from the complex and diverse data can these data be used efficiently and can be utilized to create value. Regarding how to mine this information, some countries with early development of the information technology industry have relatively complete learning mechanisms [5–7]. The two main algorithms are machine learning algorithms and deep learning algorithms. Based on such learning algorithms, it can efficiently solve various practical issues, which play a very important role in various industries and fields [8].

For insurance institutions, there is a large amount of data such as insurance policies and customer information. If the data excavated are not used in a timely and reasonable manner, resources and information will be wasted. Insurance companies should effectively mine customer information and business data based on actual needs, extract
Various insurance companies collect a large amount of business data information, but most insurance companies in China currently lack the awareness of mining potential information, resulting in a lot of invalid information, which occupies a large amount of data space [11]. Using data mining technology and machine learning algorithms, data can be effectively analyzed, the intrinsic relationship between business and data can be obtained, and high-value potential customers can be found accordingly, so as to carry out precise and scientific marketing for them and improve the insurance business [12, 13].

Precise positioning of high-value customers and promotion groups of insurance companies is the key to insurance marketing strategies. In real business, there may be orders of magnitude difference in the number of high-value and low-value customers, and the impact of misidentifying the two types of customers is also different [14, 15]. In the process of customer positioning, if customers who create lower profits for the company are identified as high-value customers, it may lead to the consequences of occupying the company's promotion resources, wasting unnecessary time and labor costs [16]. Carefully omitting some high-value customers or misjudging them as low-value customers will cause the company lose better customer promotion opportunities, which will be detrimental to the efficient and stable operation of the company [17].

Foreign scholars have studied computer technology earlier, and the research on data mining technology in China's academic circles is also based on foreign academic papers first [18, 19]. At present, the application scope of data mining technology is getting wider and wider, and with the rise of a new generation of the information technology revolution, the technology is becoming more mature. In the insurance field, many scholars also have rich experience in data collection, database establishment, and data processing [20]. As early as 2000, some foreign scholars have used a variety of data mining methods to explore the two major issues in the insurance industry. One is to analyze whether the policyholders renew their contracts, and the other is to effectively identify high-risk policyholders. Conducting a real-world case study can have an impact on how premiums are priced, which in turn affects insurers' revenues. Some Korean scholars used the data of Korean medical insurance companies as samples to predict hypertension and established three models of logistic regression, CH1AD, and C5.0 decision tree. The comparison results came to different conclusions from previous studies: CH1AD has the best predictive ability, followed by logistic regression, and C5.0 had the worst prediction effect, which can provide policy measures for hypertension management [21].

Compared with foreign research in related fields, the literature research on insurance industry data mining technology in China started relatively late. By consulting the literature, it is found that the main methods involved in data mining technology in the insurance industry include association rules, logistic regression models, decision tree models, and BP neural networks. Some scholars have mined association rules from policy data and mined the characteristics of policyholders who are prone to claim cases [22]. In order to study the reasons for customer churn, some scholars have established early warning analysis models, and based on the basic information of customers, they have used decision tree analysis methods to research and predict problems. Some scholars have visited nearly ten insurance companies [23]. Based on the collected insurance customer information, they conducted data mining and analysis with the help of a decision tree model. They mainly discussed the influencing factors behind customer compensation amounts and provided a theory for companies to determine new compensation rates [24].

This paper mainly studies the purchase choices of financial insurance based on different customers and conducts a detailed analysis of the factors that affect residents' purchase of endowment insurance in order to accurately locate the customer group according to the analysis results and provide references and ideas for academic research in this field. The customer identification model established by algorithm technology greatly improves the efficiency of customer identification and provides a more scientific method for the original simple indiscriminate marketing, which not only helps sales personnel to accurately identify high-value customers but also effectively improves competitiveness of the insurance company business [25–27].

2. Methods and Theory

2.1. Research Object. The data source of this article is the survey results of the China General Social Survey (CGSS) 2020 questionnaire (resident questionnaire). Since there are many factors affecting the purchase behavior of financial insurance, the two most important aspects are individuals and families. The questionnaire survey content involved is wide, including personal factors and family factors, which can fully select the influencing variables. The survey questionnaire is carried out nationwide, and the data distribution in the region is relatively balanced, ensuring that the final result obtained is representative.

Using SPSS software to export the data to Excel format, a total of 15,000 initial data were obtained. By reading the literature about residents' purchase of financial insurance, combined with questionnaires, 15 characteristic variables related to individuals and family-related characteristic variables are selected, which are personal factors: age, gender, education level, marital status, enterprise attributes, social security status (whether to participate in basic medical insurance; whether to participate in basic pension insurance), personal total income last year, physical health status, political outlook, and nature of work, and family factors: number of children, total family income last year, the properties you own (including co-ownership), whether you own a family car, and whether you are engaged in investment activities.

We perform data preprocessing on the original sample, including screening of variables and samples. First, the samples that did not have a clear answer to buy or not to buy financial insurance were eliminated, then the variables with
high missing values were screened, and then the samples were rescreened according to the answers of each variable; that is, the answers were vague or inaccurate deleted. For the participation of basic medical insurance and basic pension insurance and the total income of families and individuals in the last year, the samples of “not applicable,” “do not know,” and “refused to answer” are discarded. For cars, we discard the samples of “do not know” and “refuse to answer.” For the number of children, we discard the samples of “refused to answer.” For the number of houses currently owned, we discard the samples of “refused to answer,” and the answer is an empty sample of values. The models used in the demonstration can handle both discrete variables and continuous variables during fitting. Next, we will assign values to discrete features, as shown in Table 1.

The Pearson correlation test is performed on the 13 variables remaining after the initial screening and the target variable, and the obtained correlation levels and significance are shown in Table 2. According to the size of the $P$ value, its significance can be judged. It can be seen from the table that there are 3 variables that have not passed the significance test, namely, gender, marital status, and whether to participate in basic medical insurance, so these three variables will no longer be used later. To sum up, through the preliminary screening of the features and the Pearson correlation test, 10 features were obtained for subsequent modeling, including five personal factors, namely, age, education level, whether to participate in basic medical insurance, personal total income last year, political outlook, and five family factors, namely, the number of children in the family, the family last year and the whole year total income, how many properties you own, whether you own a family car, and whether you are engaged in investment activities.

The total sample size after preprocessing is 12000, of which the sample size of purchasing financial insurance is 1000, and the sample size for not purchasing financial insurance is 11000. The ratio of the sample size is about 1:12. During model training and learning, the total samples are divided into a training set and test set by 7:3, wherein 7 are training set samples and 3 are test set samples.

2.2. Research Method. As a commonly used method in data mining algorithms, decision tree has a good prediction effect on classification problems. It is mainly divided into two stages, and the specific implementation process is as follows:

1. Training stage: the training set is divided according to certain rules and divided into subsets, and the division attributes with a certain attribute as the node will be obtained and then continue to divide these subsets with the same rules and repeat this process until each subset has the same category.

2. Test phase: classify the test samples using the division rules obtained in the previous phase. First, the root node is discriminated, and the sample is classified as a child node according to the result of the attribute test. The same recursion is executed until the sample is assigned to a leaf node, and the sample belongs to the category of the current leaf node.

2.2.1. Feature Selection. Different decision tree types have corresponding criteria for selecting the optimal division attributes. Generally, there are three methods: the method based on information gain, the method based on information gain rate, and the method based on the Gini index. (1) Information Gain. The information descendant is a typical measure of sample purity, and the smaller the value of the information descendant, the higher the sample purity. For the decision tree of the binary classification problem, assuming that the proportions of positive and negative samples in the total sample set $D$ are $P_1$ and $P_2$, respectively, the information entropy of the sample set $D$ is defined as follows:

$$\text{Ent}(D) = -(p_1 \log_2 p_1 + p_2 \log_2 p_2).$$

Information gain is defined using the concept of information rot above. Assuming the total attribute set 4, there are $V$ possible discrete attribute values as one of the elements. When it is used to divide the total set $D$, 7 branch nodes will be generated, and the value of each node can be obtained, containing the information entropy of the sample. For example, the $v_{th}$ node contains all samples with the attribute value $a$ and its information correction Ent ($D^v$). However, the number of samples contained in each node of $a$ is different; that is to say, the importance of each node is different, so each node needs to be weighted, and the weight value is the proportion of the number of samples of each node. Therefore, the calculation formula of the information gain for dividing the sample set $D$ by attribute $a$ is as follows:

$$\text{Gain}(D, a) = \text{Ent}(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|} \text{Ent}(D^v).$$

Generally speaking, there is a positive correlation between the information gain value and the improvement value of purity. Therefore, the principle of maximum information gain is used when selecting the optimal attribute, that is,

$$a^* = \arg\max \text{Gain}(D, a), \ a \in A.$$  

(2) Gain Rate. When using the information gain criterion to select the optimal attribute, an obvious disadvantage is that it is inclined to the attributes with many values, which will interfere with the selection process. Quinlan proposed the use of gain for improvement in 1993, and the gain rate is defined as follows:

$$\text{Gain}_{\text{ratio}}(D, a) = \frac{\text{Gain}(D, a)}{IV(a)},$$

where $IV(a)$ is called the intrinsic value of attribute $a$, and the calculation formula is as follows:
(3) Gini Index. The measurement of sample purity by the Gini index is not based on information correction but uses the Gini value, which represents the probability that two samples randomly selected from the sample set \( D \) belong to different categories. The Gini value is negatively correlated with sample purity, and the calculation formula is as follows:

\[
Gini(D) = 1 - p_1^2 - p_2^2.
\]  

Therefore, when the Gini index is used to divide the set, the principle of minimum Gini index is used, that is,

\[
a_* = \text{argmax}Gini_{\text{index}}(D,a), \quad a \in A.
\]  

2.2.2. Pruning Strategy. The method used to solve the overfitting of the decision tree is called the pruning strategy, which can be divided into prepruning and postpruning according to pruning. Prepruning is a constraint that occurs when a tree is generated, such as when the tree reaches a certain depth, the number of samples contained under a node is less than a specified threshold, and branching to the current node will not improve the information gain. Postpruning is to take a certain method to prune after the tree is generated. There are three main methods of postpruning.

3. Results and Discussion

We set the same misclassification cost and prior probability for the positive and negative classes, set the maximum tree depth to 5, perform model fitting of the CART decision tree, and obtain the evaluation index values of the CART decision tree (Table 3). It can be seen from the table that the overall classification accuracy rates of the training set and the test set are 74.2% and 73.6%, respectively. The training set is slightly better than the test set, but there is a certain gap between the classification accuracy rates of the two sample sets for different categories. The classification effect of negative samples

<table>
<thead>
<tr>
<th>Variable name</th>
<th>Assign</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>1 = male, 0 = female</td>
</tr>
<tr>
<td>Marital status</td>
<td>1 = married, 0 = other</td>
</tr>
<tr>
<td>Whether to participate in basic medical insurance</td>
<td>1 = Yes, 0 = No</td>
</tr>
<tr>
<td>Whether to participate in the basic pension insurance</td>
<td>1 = Yes, 0 = No</td>
</tr>
<tr>
<td>Do you own a family car?</td>
<td>1 = Yes, 0 = No, 2 = no</td>
</tr>
<tr>
<td>Whether to engage in investment activities</td>
<td>1 two yes, 0 = no</td>
</tr>
<tr>
<td>Political status</td>
<td>1 = the masses, 2 = members of the communist youth league, 3 = democratic parties, and 4 = members of the communist party of China</td>
</tr>
<tr>
<td>Education</td>
<td>1 = junior high school and below, 2 = vocational high school, ordinary high school, technical secondary school, technical school, 3 = college (adult, formal higher education), undergraduate (adult, formal higher education), and 4 = Graduate and above</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>Correlation coefficient</th>
<th>p</th>
<th>Is it significant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>-0.098</td>
<td>0.000</td>
<td>Significant</td>
</tr>
<tr>
<td>Gender</td>
<td>-0.002</td>
<td>0.780</td>
<td>Not obvious</td>
</tr>
<tr>
<td>Education level</td>
<td>0.179</td>
<td>0.000</td>
<td>Significant</td>
</tr>
<tr>
<td>Marital status</td>
<td>-0.015</td>
<td>0.05</td>
<td>Not obvious</td>
</tr>
<tr>
<td>Whether to participate in basic medical insurance</td>
<td>-0.011</td>
<td>0.29</td>
<td>Not obvious</td>
</tr>
<tr>
<td>Whether to participate in the basic pension insurance</td>
<td>0.048</td>
<td>0.000</td>
<td>Significant</td>
</tr>
<tr>
<td>Personal gross income last year</td>
<td>0.045</td>
<td>0.000</td>
<td>Significant</td>
</tr>
<tr>
<td>Political status</td>
<td>0.057</td>
<td>0.000</td>
<td>Significant</td>
</tr>
<tr>
<td>Number of children in the family</td>
<td>-0.089</td>
<td>0.000</td>
<td>Significant</td>
</tr>
<tr>
<td>Total household income for the last year</td>
<td>0.081</td>
<td>0.000</td>
<td>Significant</td>
</tr>
<tr>
<td>Number of existing properties</td>
<td>0.087</td>
<td>0.000</td>
<td>Significant</td>
</tr>
<tr>
<td>Do you own a family car?</td>
<td>0.168</td>
<td>0.000</td>
<td>Significant</td>
</tr>
<tr>
<td>Whether to engage in investment activities</td>
<td>0.189</td>
<td>0.000</td>
<td>Significant</td>
</tr>
</tbody>
</table>
is not as good as that of positive samples. From a numerical point of view, the classification accuracy TPR of positive samples is nearly ten percentage points higher than that of negative samples. At the same time, combined with the ROC curve of the CART decision tree and the AUC value in Table 3, it shows that the performance of this model in the two data sets is equivalent, and the classification effect is good.

In addition, if different misclassification costs are set for positive and negative samples and then the CART decision tree model is established, it can be obtained from the classification results that the classification accuracy of the model for negative samples has been significantly improved, but for positive samples, the recognition accuracy rate of the samples is greatly reduced, resulting in an unsatisfactory overall classification effect. Therefore, the optimal model of the CART decision tree in this problem is the CART decision tree with a depth of 4 without the cost of misclassification.

As can be seen from Figure 1, the importance of the classification features of the CART decision tree on whether to purchase financial insurance is ranked from high to low as family annual income, age, personal annual income, whether to engage in investment activities, the number of children, and the level of education. Among them, the two variables of annual household income and age account for more than 40%.

From Figure 2, there are 4 characteristic variables used to establish the CART tree model, namely, family annual income, age, personal annual income, and whether to engage in investment activities. At the same time, the classification rules of the CART decision tree can be obtained: if the annual family income is less than or equal to 52000 yuan, it can be judged that the person will not buy financial insurance. If the annual family income is greater than 52000 yuan and the age is greater than 66 years old, it can be judged that the person does not buy financial insurance. If the age is less than or equal to 70 years old, but the personal annual income is greater than 40000 yuan, the model will determine that financial endowment insurance will be purchased. For residents with an annual family income greater than 52000 yuan, age less than or equal to 66 years old, and personal annual income of less than or equal to 40000 yuan, if he/she engaged in investment activities, it is determined that the person purchases financial insurance; otherwise, it is determined that it will not purchase financial insurance.

---

<table>
<thead>
<tr>
<th>Sample</th>
<th>ACC (%)</th>
<th>TPR (%)</th>
<th>TNR (%)</th>
<th>Gmean (%)</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training set</td>
<td>74.20</td>
<td>74.30</td>
<td>65.20</td>
<td>69.50</td>
<td>0.724</td>
</tr>
<tr>
<td>Test set</td>
<td>73.60</td>
<td>74.00</td>
<td>65.00</td>
<td>69.30</td>
<td>0.718</td>
</tr>
</tbody>
</table>

**Table 3: Values of each evaluation index of the CART decision tree.**

---

**Figure 1: Importance of variables in CART decision tree.**

**Figure 2: Simple tree graph of the CART decision tree.**
4. Conclusion

(1) This chapter introduces the decision tree machine learning algorithm, including the decision tree generation process, optimal attribute selection rules, and pruning strategies. The decision tree obtains the importance of the feature variables and comprehensively considers them, among which the annual family income of the family factor, the number of children, whether there is a family car, the age of the personal factor, personal annual income, education level, and whether to engage in investment activities or not are the most critical variables. Insurance institutions can pay more attention to customer information in these areas.

(2) The classification rules of the decision tree provide a reference direction for the sales business of insurance companies. Insurance marketers should focus on residents with higher household income and personal income as key marketing targets. In addition, residents who often engage in financial investment activities have high investment awareness. They should also be targeted for marketing.

Data Availability

The figures and tables used to support the findings of this study are included in the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

