Avrupa Bilim ve Teknoloji Dergisi
Özel Sayı, S. 7-11, Ocak 2018
© Telif hakkı EJOSAT'a aittir
Konferans Bildirisi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
Special Issue, pp. 7-11, January 2018
Copyright © 2017 EJOSAT
Conference Paper

An Application of Data Mining in Individual Pension Savings and Investment System

Zeynep CEYLAN[1,2]*, Samet GÜRSEV[2], Serol BULKAN[2]

[1] Corresponding author: Ondokuz Mayıs University, Faculty of Engineering, Department of Industrial Engineering, Samsun, Turkey
[2] Marmara University, Faculty of Engineering, Department of Industrial Engineering, Istanbul, Turkey,

(First received 3 November 2016 and in final form 22 July 2017)

Abstract

Individual Pension System (IPS) is a personal future investment system that allows individuals to regularly save for their retirement. IPS is enacted by the law and supported by the government through state contribution. In Turkey, IPS entered into force on October 27, 2003 and it achieved an impressive progress over the last years. This improvement has caused increase in amount of raw data stored in databases. However, accumulated data are complicated and big to be processed and cannot be analyzed by classical methods. Data mining is becoming an essential tool to discover hidden and potentially useful knowledge from raw data. For this reason, application of data mining techniques on Individual Pension Savings and Investment system is necessary. In this study, one of the data mining techniques, decision tree classification, was used to determine customers' profile. SPSS Clementine 12.0 software was used to develop a classification model. Analyses were performed by various decision tree algorithms. Some customer information of a pension company operating in Turkey were extracted from system. The significant rules about customers were revealed by analysis. The results of analysis indicated that the CHAID algorithm showed the best prediction with an accuracy of 85.64% among C5.0, C&R Tree, QUEST.

Keywords: Data mining, customer profile, individual pension system

* Corresponding author : Ondokuz Mayıs University, Faculty of Engineering, Department of Industrial Engineering , Atakum, 55139, Samsun, Turkey, zeynep.dokumaci@omu.edu.tr, phone: +90 362 312 19 19 - 1098

1. Introduction

The individual (private) pension system (IPS) is a personal future investment system which was created to complement the existing public social security systems in Turkey, such as SSK, Bağ-Kur or Government Retirement Fund. The system's target is to contribute economic development by providing long term funds and thereby increasing employment. It helps to citizens to live a more comfortable life in their retirement and to maintain their living standards. The government pays a 25% portion as government contribution for each contribution amount paid by participants. Turkish Individual Pension System started on October 27, 2003 with the contribution of six pension companies. The demographic distribution indicates most part of participants is located in Istanbul. In addition, between 25-40 ages are the most effective part of pension users and majority of this population's education level is university or master degree.

In this study, IPS customers of a well-known company in Turkey were investigated. The study aims to classify IPS customers according to their demographic attributes. This allows company to identify the profitable or unprofitable customer's profiles in detail. Thus, the company could carefully evaluate a number of important criteria that affect the potential participation of customers to the system. For example, what kind of customers would likely use these services? How many new customers could be attracted to the system? How much contribution rate they attend? and etc.

2. Research Method

2.1 Data mining

Companies databases contain hidden information. Data mining (DM), also called knowledge discovery in databases, is used for discovering interesting and useful (previously unknown and valuable) information or patterns from large data in databases. It is widely used for decision making in business (insurance, banking, retail), science research (medicine, astronomy), and government security (detection of criminals and terrorists). It is also a powerful technique with great potential to help companies focus on important information about their customers. Therefore, it provides companies to determine the impact on sales, customer satisfaction, and corporate profits (Tan et al., 2006; Chen et al., 2006; Yin et al., 2011).

In business environments, determination of profitable costumers is a significant and competitive factor of an organization. Because identification of customers to meet their product or service provides benefits to remain in the current economic environment. Moreover, analyzing customer profile helps organizations to know customers' value and special features, to identify the most profitable ones and to increase high quality relations with them. This helps much deeper and more comprehensive understanding of costumers. However, its application on individual pension savings and investment system is rare in literature.

In this study, we aimed to fill the research gap. For this purpose, decision tree model, data mining approach was performed to discover relation between demographic attributes of customers (Apeh et al., 2014).

2.2 Decision tree

Decision tree (DT) is a data mining method that is often used for classification, prediction, interpretation, and data manipulation. Because of easy interpretation and comprehension, decision tree has the advantages for the decision makers. A decision tree is a decision support tool that includes a root node, branches, and leaf nodes. Each internal node represents a test on an attribute, each branch represents an outcome of the test, and each leaf node (or terminal node) represents a class label. The top most node in a tree is the root node. The decisions are made at each node and each records of data continues through the tree along a path until the record reaches a leaf or terminal node of the tree (Han et al., 2012; Lee et al., 2007).

Several statistical algorithms such as Classification and Regression Trees (C&RT), Chi-Squared Automatic Interaction Detection (CHAID), Quick, Unbiased, Efficient, Statistical Tree (QUEST), and C5.0 (upgraded version of C4.5) have been developed for building decision trees. They work differently in the following ways. For example, how many splits are allowed at each level of the tree, when the tree is built how those splits are chosen, and how the tree growth is limited to avoid from over-fitting (Linoff and Berry, 2011; Prasad and Madhavi, 2012). These algorithms can be separated in terms of different characteristics as shown on Table 1 (Chien and Chen, 2008).

3. Application

3.1 Data description

The real data was derived from a company database in Turkey. This dataset contains 171625 customer records with 7 attributes for the year of 2014. The decision tree models generated from the dataset using demographic variables of customers. Detailed information about attributes in the customer dataset is shown in Table 2. As seen from table, the dataset includes continuous and discrete attributes. In order to measure performance of the algorithms without any bias, continuous variables was converted into categorical variables.

Table 1. Comparison of CHAID, QUEST, C5.0, and CART

Algorithm	**Data Type**	**Split Criteria**	**Number of branches at each node**	**Target variable**	**Input variable**
CHAID	Discrete	Chi-square test	Two or more	Categorical	Categorical/ Continuous
QUEST	Discrete	Chi-square for categorical variables / J-way ANOVA for continuous/ordinal variables	Two	Categorical	Categorical/ Continuous
C5.0	Discrete/ Continuous	Gain Ratio Entropy info	Two or more	Categorical/ Continuous	Categorical/ Continuous
C&RT	Discrete/ Continuous	Gini index entropy	Two	Categorical/ Continuous	Categorical/ Continuous

Table 2. Details of customer data

Variable	Variable Type	Description
Contribution rate (Target)	Categorical	Monthly contribution rate (₺) 0-250, 251-500, 501-750, 751-1000, or 1000+
Payment Type (Input)	Categorical	Credit card, Remittance, or Salary
Gender (Input)	Binary	Male or Female
Marriage Status (Input)	Binary	Single or Married
Kids Status (Input)	Binary	Yes or No
Age (Input)	Continuous	Age of customers
Education Level (Input)	Categorical	High School level and below, Bachelor degree, or Master and above

3.2 Modeling

Clementine V12.0 software developed by SPSS was used to build predictive model in classifying the customer dataset above. SPSS Clementine includes a wide variety data mining techniques associated with data preparation, manipulation and visualizations tools. In the study, the decision tree models were worked using data set node, predictive model node and assessment node.

Several algorithms (QUEST, CHAID, C5.0, C&RT) in SPSS Clementine software were employed to build the decision trees, and then all models were evaluated by using target variable (contribution rate) and input variables (payment type, age, gender, marriage status, kids status, education level). The data stream can be found in Fig. 1.

4. Results and Discussion

For the analysis, the data is firstly divided into training data and testing data. The training set is used to build the classifier, while the test set is used to validate it. In this study the percentages used for training and testing data are 75% and 25%, respectively. The accuracy measure, which is used widely to compare the performance of classifiers, was obtained by using Eq. (1).

$$Accuracy\ (\%) = \frac{TP+TN}{TP+TN+FP+FN} \qquad (1)$$

(True Negative) is the number of true negative cases and *FN (True Positive)* is the number of false negative cases. The where *TP (True Positive)* is the number of true positive cases, *FP (True Positive)* is the number of false positive cases, *TN* performance of all models was recorded by using highest testing and validation predictive accuracy rate.

As shown on Fig. 2, the CHAID decision tree model is more accurate than other techniques. In overall performance, the CHAID model recorded an impressive of accuracy score of 85.64% compared of that to the other models: C5.0 (84.59%), C&RT (82.63%) and QUEST (64.35%), see Fig. 3.

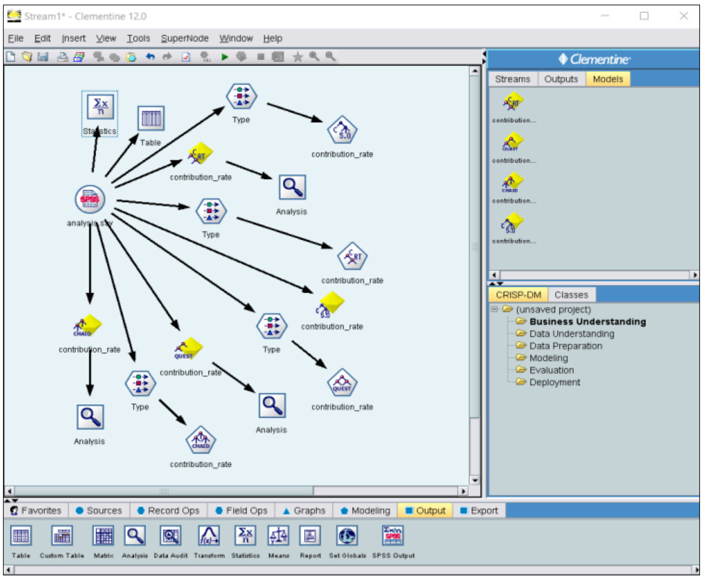
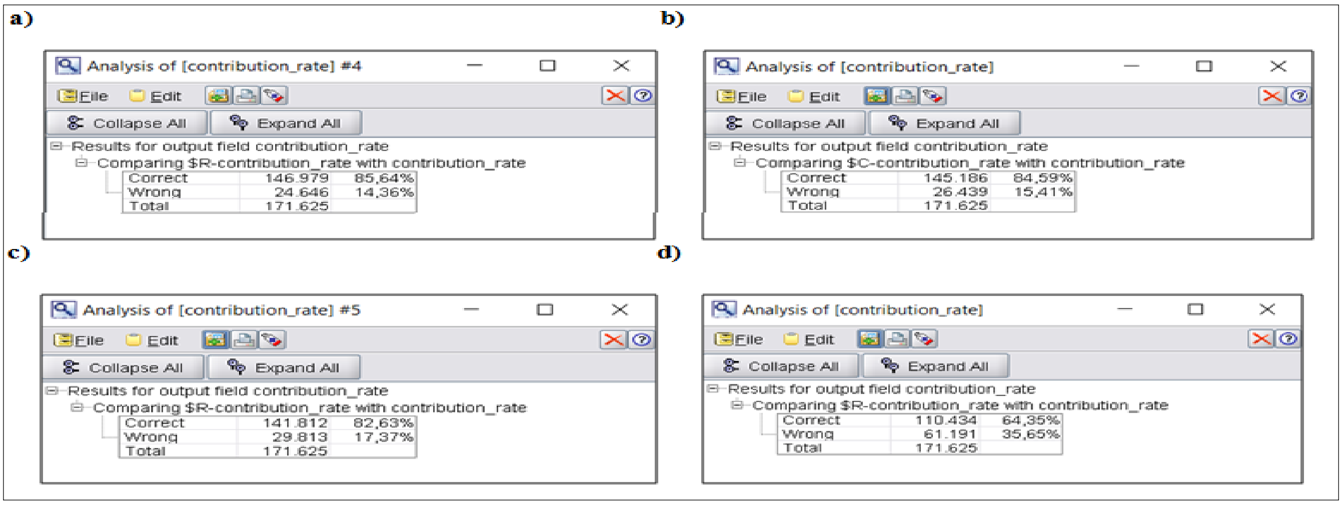

Stream1* - Clementine 12.0
File Edit Insert View Tools SuperNode Window Help
Statistics
Table
Type
contribution_rate
Analysis
analysis.sav
Clementine
Streams Outputs Models
contribution...
CRISP-DM Classes
(unsaved project)
Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Favorites Sources Record Ops Field Ops Graphs Modeling Output Export
Table Custom Table Matrix Analysis Data Audit Transform Statistics Means Report Set Globals SPSS Output

Figure 1. SPSS Clementine workspace with four decision tree models


Analysis of [contribution_rate] #4
File Edit
Collapse All Expand All
Analysis of [contribution_rate]
Analysis of [contribution_rate] #5

a)

```
⊟ Results for output field contribution_rate
   ⊟ Comparing $R-contribution_rate with contribution_rate
      Correct   146.979   85,64%
      Wrong      24.646   14,36%
      Total     171.625
```

b)

```
⊟ Results for output field contribution_rate
   ⊟ Comparing $C-contribution_rate with contribution_rate
      Correct   145.186   84,59%
      Wrong      26.439   15,41%
      Total     171.625
```

c)

```
⊟ Results for output field contribution_rate
   ⊟ Comparing $R-contribution_rate with contribution_rate
      Correct   141.812   82,63%
      Wrong      29.813   17,37%
      Total     171.625
```

d)

```
⊟ Results for output field contribution_rate
   ⊟ Comparing $R-contribution_rate with contribution_rate
      Correct   110.434   64,35%
      Wrong      61.191   35,65%
      Total     171.625
```

Figure 2. a) Model performance for CHAID algorithm b) Model performance for C5.0 algorithm c) Model performance for C&RT algorithm d) Model performance for QUEST algorithm

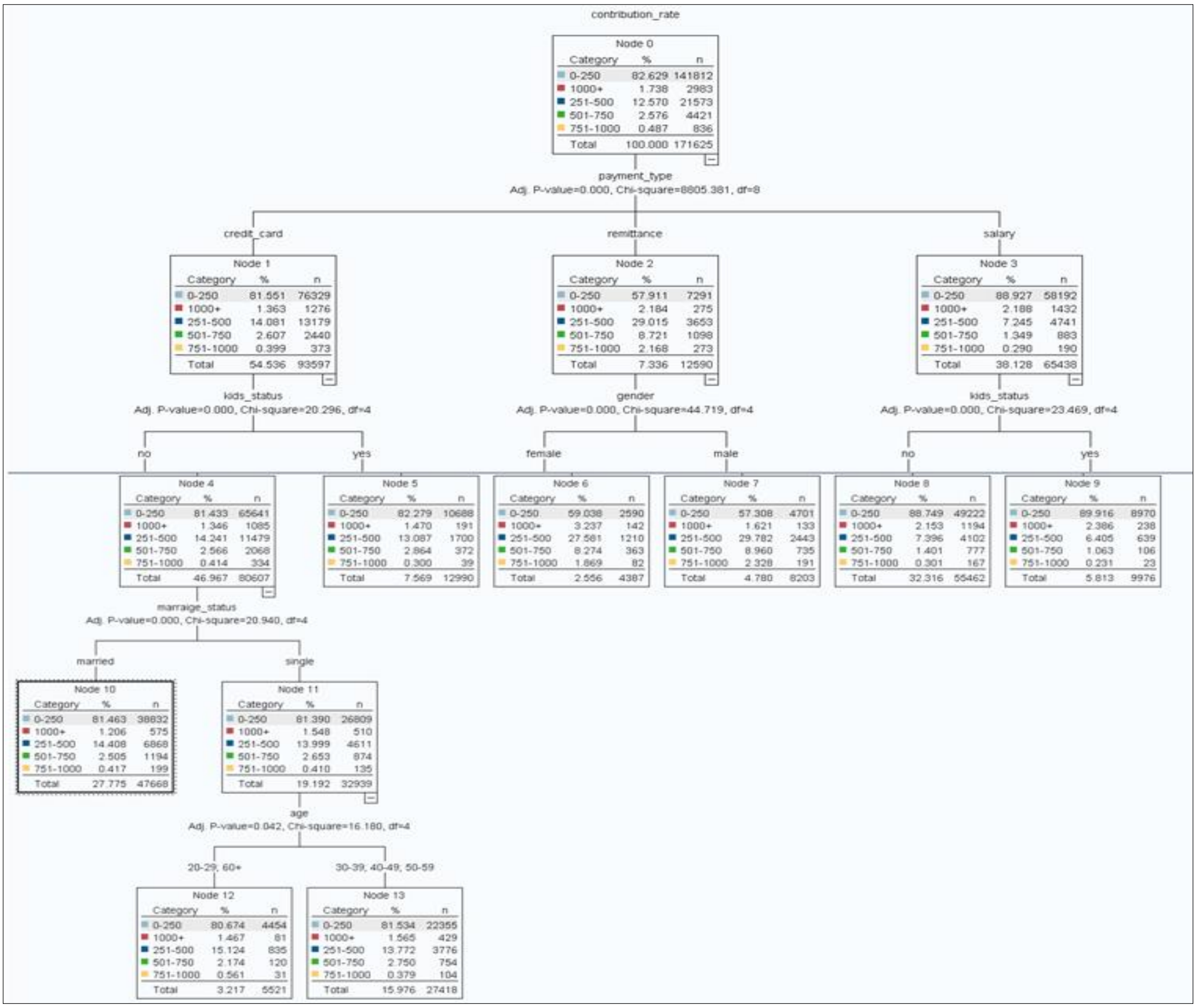


Figure 3. A sample part of CHAID classification tree

5. Conclusion

Determination of customers profile is very important for sales and marketing strategies for companies. For this reason, analyzing of customer behavior according to their features will be necessary. Data mining provides the companies to define and classify customers in detail to understand their behaviors. In addition, data mining provide strategic information for many customer-centric applications.

Data mining focuses on the practical importance of the information or knowledge gained from the models. The study results showed that the construction and evaluation of decision tree algorithms on individual pension system is important. This study will be a guiding start for Individual Pension Savings and Investment System. It might be beneficial for further studies for application of other data mining techniques. Using customer profiling through data mining might be the new competitive strength for the company owners in this sector

References

Apeh E., Gabrys B., Schierz A. 2014. Customer profile classification: To adapt classifiers or to Relabel customer profiles?, Neurocomputing 132, 3-13.

Chen Y.-L., Chen J.-M., Tung C.-W. 2006. Data Mining Approach for Retail Knowledge Discovery with Consideration of the Effect of Shelf-Space Adjacency on Sales. Decision Support Systems 42, 1503–1520.

Chien C.-F., Chen L.F. 2008. Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry. Expert systems with applications 34, 280-290.

Han J., Kamber M., Pei J. 2012. Data Mining Concepts and Techniques. Morgan Kaufmann Publishers Inc., 3. Edition, Waltham, USA.

Lee S., Lee S., Park Y. 2007. A prediction model for success of services in e-commerce using decision tree: E-customer's attitude towards online service. Expert Systems with Applications 33, 572-581.

Linoff G.S., Berry M.J. 2011. Data mining techniques: For marketing, sales and customer relationship management. (3rd ed.) Indianapolis, Wiley Publishing Inc.

Prasad U.D., Madhavi S., 2012. Prediction of Churn Behavior Of Bank Customers Using Data Mining Tools. Business Intelligence Journal 5(1), 96.

Tan P.-N., Steinbach M., Kumar V. 2006. Introduction to Data Mining. International Edition, Pearson Education Inc., Boston, USA.

Yin Y., Kaku I., Tang J., Zhu J.M. 2011. Data Mining: Concepts, Methods and Applications in Management and Engineering Design.