International Journal of
Intelligent Systems and Applications in Engineering

ISSN:2147-6799 www.ijisae.org Original Research Paper

Knowledge Discovery on Investment Fund Transaction Histories and Socio-Demographic Characteristics for Customer Churn

Fatih CIL[1], Tahsin CETINYOKUS*[2], Hadi GOKCEN[2]

Accepted : 11/12/2018 Published: 31/12/2018

Abstract: The need of turning huge amounts of data into useful information indicates the importance of data mining. Thanks to latest improvement in information technologies, storing huge data in computer systems becomes easier. Thus, “knowledge discovery” concept becomes more important. Data mining is the process of finding hidden and unknown patterns in huge amounts of data. It has a wide application area such as marketing, banking and finance, medicine and manufacturing. One of the most commonly used application areas of data mining is recognizing customer churn. Data mining is used to obtain behavior of churned customers by analyzing their previous transactions. In the same manner using with obtained tendency, other active customers are held in the system. It is possible to make by various marketing and customer retention activities. In this paper, it is aimed to recognize the churned customers of a bank who closed their saving accounts and determine common socio-demographic characteristics of these customers.

Keywords: *Data mining, Customer churn, Decision trees and classification rules, Mutual funds*

1. Introduction

After the increase in the amount of data about customer’s transactions and escalating competition in the industries, firms have attempted to use their existing data efficiently in order to obtain information which is used to understand purchasing and service characteristics of customers. Due to increasing competition and improvements in IT Technologies, obtained information will provide advantage to firms, so it has been gaining importance.

The increasing usage of database systems and their extraordinary increase in volume have made firms to be faced with the issue how firms will derive benefit from these data. Traditional query or reporting tools are inadequate in the face of massive amounts of data. This is why the new tendencies and Data Mining is a result of this search [1].

In this study, it is focused on two objectives. The first of these is to learn the details of the transactions of the customers buying and selling bank mutual funds including former customers who closed their bank accounts after a stated transaction history. The second objective is to obtain socio-demographic characteristics of the customers who closed their investment account. With obtained results, the rules are determined to prevent losing customers who are tended to close their accounts. The main contribution is the first study in literature considering mutual funds customer churn based on data mining techniques.

It is organized as follows. In the second section, data mining and classification algorithms used in the study are discussed. Then, a general overview to bank mutual funds is presented in the third section. In the fourth section, the study conducted is mentioned and finally the results of the study are covered.

2. Data Mining, Decision Trees and Classification Rules

2.1. Data Mining

Data mining is a set of techniques that enable the extraction of large volume data sets useful results. In problem solving, located data in different sources or databases are analyzed. As a result of this analysis, hidden images will emerge. Using the obtained patterns can be given strategic management decisions.

Data mining is discovery driven. It involves various techniques including statistics, decision trees, neural networks, genetic algorithms and visualization techniques. Data mining has been applied in many fields such as marketing, finance, banking, health care, customer relationship management and organizational learning [2].

Owing to the advancement of information technologies and global competitiveness, data mining has increasingly become an important and a widely known field. Data mining applications has been varied recently, although exists in academic framework for long years [3].

The major reason that data mining has attracted a great deal of attention in the information industry in recent years is due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge. The information and knowledge gained can be used for applications ranging from business management, production control and market analysis to engineering design and science exploration [1].

Data mining is searching of relations and rules for future prediction by using software. Assuming that near future will not be very different from today or past; rules obtained from existent data will help us to form correct estimates of future trends.

Knowledge discovery in databases (KDD) is processing of data by data mining techniques in databases. Processing large amounts of data in data warehouses can be possible with new generation tools and techniques. So, studies applied about KDD keep up to date.

[1] Operations Business Development at Finansbank İstanbul, TURKEY
[2] Industrial Eng., Gazi University, Ankara – 06570, TURKEY
** Corresponding Author: Email: tahsinc@gazi.edu.tr*

According to some sources; knowledge discovery in databases refers to the overall process of discovering useful knowledge from data, and data mining refers to a particular step in this process [4].

2.2. Decision Trees and Classification Rules

Classification data mining techniques are used when the dependent (prediction) variable is discrete. Inferential techniques are used to create decision trees and rules that effectively classify the dataset as a function of its characteristics. Decision tree solution methods provide automated techniques for discovering how to effect such classifications [5,6]. At the top of the tree there is a root node, at which the classification process begins. The test at the root node tests all instances, which pass to the left if true, or to the right if false. At each of the resulting nodes, further tests serve to continue classifying the data. Each leaf of the tree represents a classification rule. Rule-based solutions translate the tree by forming a conjunct of every test that occurs on a path between the root node and the leaf node of the tree [5]. One of the advantages offered by rule induction algorithms is that the results may be directly inspected to understand the variables that can be effectively used to classify the data. Variables at the root node represent the strongest classifiers (denoted as level 1 in the results section), followed by the next strongest classifiers at each of the leaf nodes (denoted as level 2, 3, etc.).

In this study Id3 and J4.8 decision tree algorithm with PART, JRip ve OneR classification rules are used. Id3 (basic divide-and-conquer decision tree algorithm), J4.8 (implementation of C4.5 decision tree learner), PART (obtain rules from partial decision trees built using J4.8), JRip (RIPPER algorithm for fast, effective rule induction), OneR (1R classifier) (Witten and Frank, 2005).

3. Bank Investment Funds

The accumulation of all kinds of investors, whether large or small, is managed by professional managers by distributing to the various capital market instruments. These assets are called funds. Fund portfolios consisting of treasury bonds, government securities, repo, common stocks and other capital market instruments are managed by professional portfolio managers in accordance with internal regulations.

According to Capital Market Regulations investment fund portfolios must be managed by institutions and portfolio management companies. Investment funds are audited periodically by Capital Markets Board and independent external audit firms.

Fund prices are determined by fund founder and these determined prices are used next day in buying and selling funds.

Fund portfolio value is calculated in accordance with stock market prices that funds are bought and sold. Total fund value is divided by total share number that circulates on the day of valuation and fund price is calculated.

Interests, dividends, trading gains and daily gains of portfolio assets are recorded as income to fund at the same day and reflected to the fund prices. Therefore, if an investor sells his/her funds, he/she gets his/her share from the fund gaining

3.1. Discovery Studies on Investment Fund and Customer Churn

Data mining are often used in studies about marketing, finance, banking, manufacturing, health care, customer relations management and organizational learning. Some of these studies have been interested in the retaining the customers and minimizing the customer churn.

Masand et al. [8] developed an automated system used historical data of cellular phone customers. The system periodically identifies churn factors for several statuses. In this study, decision trees algorithms and neural networks algorithms were used.

Poel and Larivière [9] investigate predictors of churn incidence for financial services as part of customer relationship management (CRM) in their stuies. Their findings suggest that: demographic characteristics, environmental changes and stimulating ‘interactive and continuous’ relationships with customers are of major concern when considering retention; customer behaviour predictors only have a limited impact on attrition in terms of total products owned as well as the interpurchase time.

Ahn et al., [9] investigated determinants of customer churn in the Korean mobile telecommunications service market. In the study, mediating effects of a customer's partial defection on the relationship between the churn determinants and total defection were analyzed and their implications were discussed. Results indicate that customer's status change explains the relationship between churn determinants and the probability of churn.

Ruta et al., [10] proposed a new k nearest sequence (kNS) algorithm along with temporal sequence fusion technique to predict the whole remaining customer data sequence path up to the churn event in telecom industry.

Chu et al., [11] proposed a hybridized architecture to deal with customer retention problems. In the system, common clustering-followed- by- classification approaches have been used.

Aydoğan et al., [12] studied in a cosmetic firm to determine the customer segment which tends to leave. So they developed customized campaigns and marketing strategies for customer profile who likely to leave. Clustering techniques were used for segmentation and classification techniques were used for determining the customer churn.

Telecommunications industry, customer churn is recognized as an important research topic. In this view, Gopal and Meher [13] have discussed the use of ordinal regression method to estimate the time of customer churn and the duration of customer retention. On the modeling of the customer churn, ordinal regression technic has used in this study for the first time.

Farquad et al., [14] proposed a hybrid algorithm to predict churn rates of customer using credit card. This algorithm is combination of Support Vector Machine method and Naive Bayes method.

Huang et al., [15] presented a new set of features for broadband internet customer churn prediction, based on Henley segments, the broadband usage, dial types, the spend of dial-up, line-information, bill and payment information, account information. The experimental results showed that the new features with using these four models (Logistic Regressions, Decision Trees, Multilayer Perceptron Neural Networks and Support Vector Machines) techniques are efficient for customer churn prediction in the broadband service field.

Chen and Fan [16] proposed a framework of distributed customer behavior prediction using multiplex data. Framework that is to adapt to the emerging distributed computing environment, a novel approach called collaborative multiple kernel support vector machine (C-MK-SVM) is developed for modeling multiplex data in a distributed manner, and the alternating direction method of multipliers (ADMM) is used for the global optimization of C-MK-SVM.

Zhang et al., [17] investigated the effects of interpersonal influence on the accuracy of customer churn predictions and proposed a novel prediction model that is based on interpersonal influence and that combined the propagation process and customers’ personalized characters.

Slavescu and Panait [18] in their studies focuses on how to better support marketing decision makers in identifying risky customers in telecom industry by using Predictive Models. Based on historical data regarding the customer base for a telecom company they proposed a Predictive Model using Logistic Regression technique and evaluate its efficiency as compared to the random selection.

Devi and Madhavi [19] prepared a modeling study on purchasing behavior of bank customers. In this study churn customers have been predicted in India.

Jamal and Tang [20] in their study are directed for methods to developing customer retention and loyalty strategies. One of them is about calculating likelihoods of next action taken by customers. Another one is calculating risk scores of customer churn based on customer attributes.

A social network analysis for customer churn prediction was improved by Verbekea et al. [21]. A new model setup was introduced in order to combine a relational and nonrelational classification model. It was shown that the model setup boosted the profits generated by a retention campaign.

Farquad et al., [22], an analytical CRM (customer relationship management) application was achieved with churn prediction using comprehensible support vector machine. It was indicated that the generated rules acted as an early warning expert system to the bank management.

A churn prediction in telecommunication industry was improved using data mining techniques by Keramati et al. [23]. Above 95% accuracy for Recall and Precision was achieved with the proposed methodology in addition, a new methodology was introduced for extracting influential features in dataset. As can be seen in the literature, there are no studies on the banking sector mutual fund clients.

4. Application

Due to increasing competition and new emerging actors in banking sector, banks' efforts for keeping customers and getting maximum benefit from them increased.

Finding new customers, advertising and promotion costs to find new customers increase day by day. Even though lost customers are compensated by new customers, cost of keeping customers is lower than cost of finding new customers.

The aim of this paper is to determine which transaction history and socio-demographic characteristics of customers cause them to close their accounts. With these findings determining the customers that tend to close their accounts is aimed, as well. The data of nearly 87.000 fund customers' annual investment behaviors and 65.525 customers' socio-demographic characteristics are considered. In the analysis, nearly 4.000.000 transaction of investment funds data that belongs to these customers is used.

4.1. Preprocessing of Socio-Demographic Characteristics Data of Customers

Socio-demographic characteristics data of 65.525 customers has been taken from a bank's database and preprocessed. The socio-demographic characteristics data that has been used and preprocessing procedure is as follows;

The living city information for customers has been taken from bank's data base as city codes. To make city codes categorical and add development value of cities to the study, classification has been made by using per capita GDP data of cities. According to Bank Investment Funds Specialists, city variable must get maximum 5 categorical values. Taking into consideration the Bank Investment Funds Specialists' views, k-means algorithm, commonly used in classification, has been applied from 2 to 6 for class number. Square errors are shown in table 1.

Table 1. Sum of square errors according to K-means algorithm (%)

Class Number	2	3	4	5	6
Sum of Square Errors	0.25691	0.25688	0.25687	0.25686	0.25686

Considering the Table 1, it is seen that sum of square errors is decreasing. Because Bank Investment Funds Specialists think that city variable must get maximum 5 categorical values, class number is determined as 5.

According to determined class number, 5, and value range for classes, cities that customers live and categorical values for that cities are shown in table 2.

Table 2. Cities and Categorical Values

Cities	Per capita GDP	Categorical Values
Kırklareli, Yalova, Muğla, İzmir, İstanbul, Zonguldak, Ankara, Kırıkkale, Bilecik, Eskişehir, Bursa, Tekirdağ, Manisa, İçel, Edirne	2339 - 4214 TL	1
Çanakkale, Antalya, Artvin, Denizli, Nevşehir, Sakarya, Aydın, Karaman, Balıkesir, Burdur, Rize, Kilis	1817 - 2338 TL	2
Kayseri, Kütahya, Kastamonu, Niğde, Hatay, Elazığ, Samsun, Çorum, Gaziantep, Karabük, K.Maraş, Tunceli, Konya, Isparta, Trabzon, Kırşehır, Sinop, Giresun, Amasya, Uşak, Malatya, Sivas, Diyarbakır, Afyon, Batman, Erzincan, Osmaniye, Düzce, Çankırı, Siirt, Gümüşhane, Ordu, Erzurum, Bartın, Bayburt, Şanlıurfa, Mardin, Aksaray, Adıyaman	920 - 1816 TL	3
Kars, Van, Iğdır, Yozgat, Ardahan, Hakkâri, Bingöl, Bitlis, Şırnak, Muş, Ağrı, Tokat	∞ - 919 TL	4
Kocaeli, Bolu	4215 - ∞ TL	5

City variable has been used as a categorical variable in the study.

Table 3 gives the description of variables after preprocessing of socio-demographic characteristics data of customers.

After 65,525 customer social- demographic data has been passed through pre-processing stage as defined above, a matrix which has 65525 x 15 elements has been acquired.

Table 3. Description of variables

Variable	Description	Range
Customer numbers	Customer numbers are begun from 1 and given to all customers by increasing one. Customer number variable has been used as numerical variable.	1- 65.525
Gender	Gender information of customers has been taken from bank's data base and changed as "1" for "Male" and "2" for "Female". Gender variable has been used as a categorical variable in the study.	"1", "2"
Age	By using the date of birth information the age of customers has been determined and identified age groups have been categorized.	younger than 20, "1" between 21 – 30, "2" between 31 – 40, "3" between 41 – 50, "4" between 41 – 50, "5" older than 60, "6"
Marital status	Marital status data has been changed as "1" for "Married" and "2" for "Single".	"1", "2"
Educational background	Educational background information for customers has been taken from bank's data base as "Primary School", "Elementary School", "Primary and Elementary School", "High School", "Vocational School", "Undergraduate", "Graduate" and "Postgraduate" and changed in preprocessing.	Primary School, "1" "Elementary School", "2" "Primary and Elementary School", "3" "High School", "4" "Vocational School", "5" "Undergraduate", "6" "Graduate", "7" "Postgraduate" "8"
City	The living city information.	"1", "2", "3", "4", "5"
House that customers live	House information for customers has been taken from bank's data base and has been changed.	"Own house", "1" "Rental", "2" "Dig", "3" "Belongs to family members", "4"
Experience in the last job	Job experience (year) information for customers.	less than 1 year, "1" between 2 – 5 years, "2" between 6 – 10 years, "3" more than 11 years, "4"
Monthly net income	Net income of customer has been filtered.	< 1.000 TL, "1" 1.001 - 1.500 TL, "2" 1.501 - 2.500 TL, "3" 2.501 - 4.000 TL, "4" 4.001 TL >, "5"
Occupation	Occupation type of customer information has been filtered.	Public Sector, "1" Private Sector, "2" Self Employed, "3" Supplementary Income Holder, "4" Retired, "5" Student, "6" Housewife, "7" Unemployed, "8"
Spouse employee	Customer's spouse employed status data has been changed as "1" for "Yes" and "2" for "No".	"1", "2"
Automobile of customer	Information that whether or not customer has got an automobile. "1" for "Yes" and "2" for "No".	"1", "2"
Capital	TL value of customer's total amount of investment fund.	
Asset / Liability	Information that whether or not the customer closed his / her investment account	Closed accounts, "P" Unclosed accounts, "A"
Liability date	The month in which customer made his / her account inactive. Liability date variable has been used in order to organize investment fund buy / sell transaction table.	"07.20**", "08.20**", "09.20**", "10.20**", "11.20**", "12.20**" and "01.20**"

4.2. Pre-processing of Investment Fund Buy / Sell Transaction Data

For selected customers, about 4,000,000 investment funds buy / sell transaction data have been filtered and listed for every single customer in terms of date. Listed transactions have been summed up in terms of months and transaction amount of every single customer's in these months have been determined and then percentage changes on principal amount for July 20** of transaction amounts have been calculated. For a sample customer, investment fund buy / sell order data is given in table 4.

Table 4. For a sample customer, investment fund buy / sell order data

Customer Number	Capital	July 20**	Aug. 20**	Sep. 20**	Oct. 20**	Nov. 20**	Dec. 20**	Jan. 20**
1	18759.72	-0.26	0	1.47	1.63	1.61	1.76	2.26

From variables in customer's social – demographic properties, by assessing "Liability Date" as "t-1" month, investment fund buy / sell transaction order data of every single customer have been reorganized.

For example, the customer having 1490 customer number made his

/ her account inactive in December 20**. So for this customer, “t-1” month is December 20**, “t-2” month is November 20**, “t-3” month is October 20**, “t-4” month is September 20**, “t-5” month is August 20** and “t-6” month is July 20**. “t-7” cell was filled as zero (0). Investment fund buy / sell order data of the customer having 1490 customer number is given in table 5.

Table 5. Investment fund buy / sell order data of the customer having 1490 customer number

Customer Number	Capital	t-7	t-6	t-5	t-4	t-3	t-2	t-1
1490	2642.78	0	-0.23	-0.02	0.31	-0.01	0	-1

In order to categorize percentages of organized investment fund buy / sell order data, series of processed has been done. In order to categorize zero values, not increased or decreased values, “S” code meaning “fixed” has been given. In order to categorize decreased percentage rates, clumping has been done.

According to opinion of bank investment fund specialists, decreasing percentage rate values could be clumped into between 7 and 10 clusters and shouldn’t be more than 10 clusters. In this process, regarding bank investment fund specialists’ opinion, k-means algorithms used commonly in clumping have been used. K-means algorithm has been implemented for number of clusters between 7-10 and table 6. error squares have been got.

Table 6. Sum of k-means algorithm decreasing percentage values’ error squares

Number of Clusters	7	8	9	10
Sum of Square Errors	78.79	73.46	59.72	55.77

When sum of error squares is analyzed, sharp decrease between state of 8 clusters and 9 clusters has been seen (look Figure 1.) For this reason, cluster number has been determined as 8.

According to determined number of clusters and cluster range values, decreasing percentage rates have been categorized. Categorizing process is built by gathering “D” meaning decrease in percentage rates and clusters number. Information showing which decreasing percentages belongs to which category is shown in table 7.

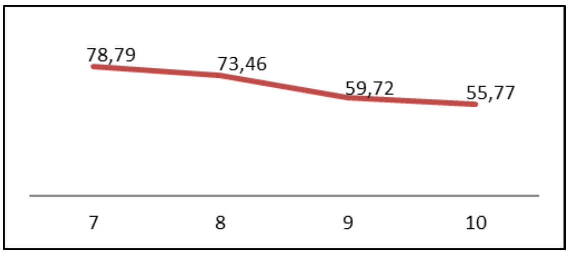

Fig 1. Change in sum of k-means error squares graph (Table 6.)

Clumping has been also made in order to categorize increasing percentage rates. Aim of this is to determine how many clusters increasing values should be classified.

According to bank investment fund specialists’ opinion, increasing percentages should be separated at most 4 clusters. Because they think that as the study has been made on the customers who closed their investment fund accounts, analyzing increasing percentages is not useful for analysis. Table 8. has been acquired by implementing k-means algorithm for between 2-4 clusters.

Table 7. Categorical classification of decreasing percentages

Categorical Values	Range of Decreasing Percentages	Number of Values
D8	(-1.00) - (-0.81)	3409 (%5)
D7	(-0.80) - (-0.54)	3318 (%5)
D6	(-0.53) - (-0.35)	4781 (%7)
D5	(-0.34) - (-0.21)	6717 (%10)
D4	(-0.20) - (-0.11)	10045 (%15)
D3	(-0.10) - (-0.05)	14133 (%21)
D2	(-0.04) - (-0.02)	15794 (%24)
D1	(-0.01) - (0)	7818 (%12)

When the table analyzed, it is clear that sum of error squares is continuously decreasing. According to bank investment fund specialists’ opinion, increasing percentages should be separated at most 4 clusters and for this reason number of clusters has been determined as 4.

According to determined number and range of clusters, increasing percentages have been categorized. Process of categorizing has been built by gathering “Y” meaning increasing percentages and clusters number. Information showing which increasing percentages belongs to which category is shown in table 9.

Table 8. Error square sum of k-means algorithm decreasing percentage values

Number of Clusters	***2***	***3***	***4***	***5***
Sum of Square Errors	0.90	0.35	0.22	0.20

Table 9. Categorical classification of increasing percentages

Categorical Values	Range of Increasing Percentages	Number of Values
Y4	∞ - (2.00)	13086 (%4)
Y3	(2.00) - (0.77)	37448 (%11)
Y2	(0.76) - (0.26)	80658 (%25)
Y1	(0.25) - (0)	196452 (%60)

Investment fund buy / sell order data for categorized percentages and customer having 1490 customer number is shown in table 10.

By gathering social demographic data from pre-processing stage and investment fund buy / sell order data, a matrix which has 65525 x 9 elements has been acquired.

Table 10. Investment fund buy / sell order data for categorized percentages and customer having 1490 customer number

Customer Number	Capital	t-7	t-6	t-5	t-4	t-3	t-2	t-1
1490	2642.78	S	D5	D2	Y2	D1	S	D8

4.3. Investment Fund Buy / Sell Transaction Order Data Analysis

By this classification, it is aimed to determine transactions during separation process by analyzing transaction details of customers who did not want to make investment fund transaction anymore and closed their accounts.

Customer investment fund buy / sell order data has been analyzed by J4.8, PART, JRip, Naive Bayes and OneR classification algorithms. In classification techniques, as object variable, “Asset

/ Liability" variable showing that customer has made his / her account inactive has been used.

Decision trees, built by analyzing investment fund buy / sell order data by classification algorithms have been analyzed. As PART, JRip and OneR algorithms are rule-learner algorithms, their results are also in rule forms. But since J4.8 gives results in the form of decision tree, the decision tree acquired by J4.8 algorithm has been converted into rule. After this conversion, 31 rules have been acquired.

Analysis results of classification algorithms Number of Rules / Number of Leaves, True Positive Rate (TP Rate), False Positive Rate (FP), Precision are analyzed by Kappa statistics and Confusion matrixes. Comparison table of analysis results have been prepared and are given in table 11.

Table 11. Comparison of classification results of investment fund buy / sell order data

Algorithm	Number of rules / Number of elements of decision tree	Number of rules for passive customers	Correctly classified percentages for passive customers	Not correctly classified percentages for passive customers	Kappa Statistics	Number of not correctly classified customers
J4.8	123	31	95.8	0.1	0.9517	62
PART	79	25	95.6	0.1	0.9516	65
JRip	14	14	99.2	0.2	0.9515	12
OneR	1	1	38.5	0.2	0.5144	914

When classification results are analyzed, it has been seen that, the highest number of rules are derived by J4.8 algorithm. Most powerful ones of these rules are;

- ***Rule 1:*** If (t-2 = D8) and (t-1 = S) = P (492, % 33.13)
- ***Rule 2:*** If (t-7 = S) and (t-6 = S) and (t-1 = D8) = P (354, % 23.84)
- ***Rule 3***: If (t-3 = D8) and (t-2 = S) and (t-1 = S) = P (212, % 14.28)
- ***Rule 4:*** If (capital > 85.95) and (t-4 = D8) and (t-3 = S) and (t-2 = S) and (t-1 = S) = P (88, % 5.93)
- ***Rule 5***: If (t-7 = S) and (t-6 = Y2) and (t-1 = D8) = P (71, % 4.78)

The rules built reveals transaction characteristics of the customers who closed their accounts. These rules will be used to determine potential customers who may be close their accounts in the future. However, these rules do not reveal which social demographic characteristics in which rules. Customer social demographic characteristics can be included by a new analysis. In order to do this, the customers obeying the rules have been determined and customer social demographic data has been reorganized. Further analysis has been made by reorganized social demographic data and classification techniques.

4.4. Customer Socio-Demographic Data Analysis

Investment account bank mutual fund customers that have ceased to be turning off customers analyzed with Id3, J4.8, PART and JRip classification algorithms. The classification techniques as the target variable that specifies the path that the customer's investment account while making passive rule number is used as a variable.

The Classification results of the analyzed algorithms were analyzed criteria such as True Positive Rate (TP Rate), False Positive, Kappa statistics and Confusion matrix. Analysis results comparison table was created. The comparison chart is given in table 12.

Table 12. The classification results comparison chart of customer socio-demographic data

Algorithm	Percentage of correctly classified records	Percentage of incorrectly classified records	Kappa Statistics	Number of incorrectly classified records
Id3	98.04	1.95	0.9757	29
J4.8	50.84	49.15	0.3244	730
PART	57.64	42.35	0.4373	629
JRip	33.53	66.46	0.008	987

When the classification results were examined, the accuracy rate of 98.04% was observed with the Id3 algorithm is the highest percentage of correctly classified records of the algorithm.

The investment fund buying selling instruction data of investment account bank mutual fund customers, who after a certain transaction history has ceased to be turning off customers by disabling their investment account and socio-demographic data classification of results of two-stage decision tree has been reached. In a bank investment fund performs but a certain transaction history after the investment account to disable the bank mutual fund customers to be out of the customers' investment fund buying selling instruction data and socio-demographic data classification has resulted two-stage decision tree. The mutual fund sales practice data involves the rules that shows after which motion the customers closed down their investment account. The socio-demographic characteristics of the customers were derived from the classification of the clients' socio-demographic characteristics.

4.5. The Combination of the Customer's Investment Account Closed Down with the Socio-Demographic Characteristics

Classification results for investment account closing transactions of the customer done by J4.8 algorithm and classification results for social – demographic characteristics of the customer have been combined. With this combination, the decision tree having two levels can be used as a decision tree having only one level. Some of the decision tree leaves with highest explanation power is given below as examples and the decision tree structure of Rule 1 is given figure 2.

Rule 1: If (t-2 = D8) and (t-1 = S) = P (492, 33.13%)
Socio-demographic characteristics;

- (Occupation = 2)
- (Occupation = 3) and (Age = 2) and (Experience (year) in the last job = 2) and (Gender = 1)
- (Occupation = 3) and (Age = 3) and (Monthly Net Income = 3) and (Educational Background= 6)
- (Occupation = 3) and (Age = 3) and (Monthly Net Income = 3) and (Educational Background= 8)
- (Occupation = 3) and (Age = 3) and (Monthly Net Income = 5) and (Marital Status= 1)
- (Occupation = 3) and (Age = 3) and (Monthly Net Income = 5) and (Marital Status= 2) and (Educational Background= 5)
- (Occupation = 3) and (Age = 4) and (Experience (year) in the last job = 1) and (Method of Study= 3) and (Gender = 1)

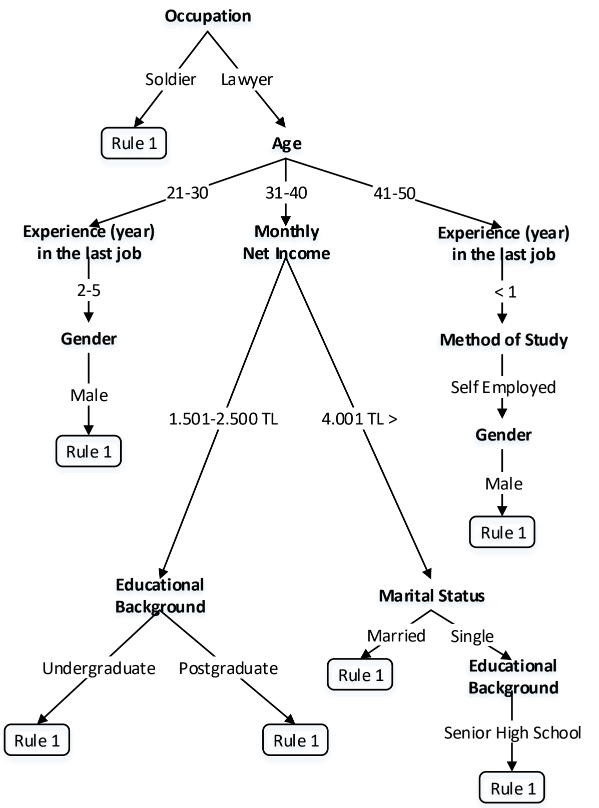

Fig 2. The decision tree structure of Rule 1

Rule 2: If (t-7 = S) and (t-6 = S) and (t-1 = D8) = P (354, % 23.84) Socio-demographic characteristics;

- (Occupation = 3) and (Age = 3) and (Monthly Net Income = 2) and (Gender = 1)
- (Occupation = 3) and (Age = 3) and (Monthly Net Income = 5) and (Marital Status= 2) and (Educational Background= 7)
- (Occupation = 3) and (Age = 5) and (Experience (year) in the last job = 2) and (Monthly Net Income = 2)
- (Occupation = 4) and (Gender = 1)
- (Occupation = 5) and (Age = 2)

The decision tree structure of Rule 2 is given figure 3.

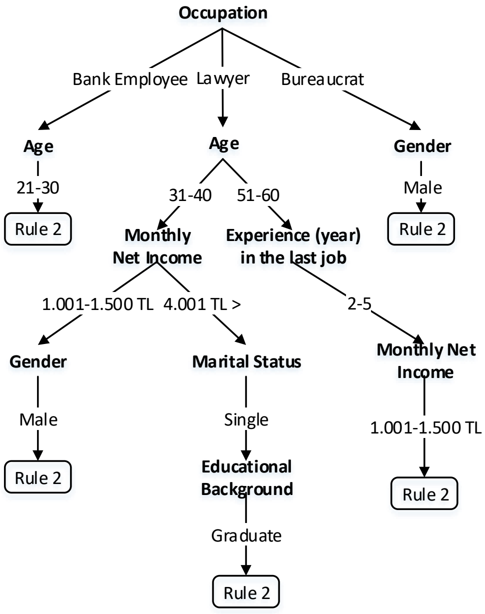

Fig 3. The decision tree structure of Rule 2

Rule 3: If (t-3 = D8) and (t-2 = S) and (t-1 = S) = P (212, 14.28%) Socio-demographic characteristics;

- (Occupation = 3) and (Age = 2) and (Experience (year) in the last job = 1) and (City = 1)
- (Occupation = 3) and (Age = 4) and (Experience (year) in the last job = 1) and (Method of Study= 2) and (Gender = 2) and (Monthly Net Income = 3)
- (Occupation = 3) and (Age = 4) and (Experience (year) in the last job = 3)
- (Occupation = 3) and (Age = 4) and (Experience (year) in the last job = 4) and (Gender = 1)

The decision tree structure of Rule 3 is given figure 4.

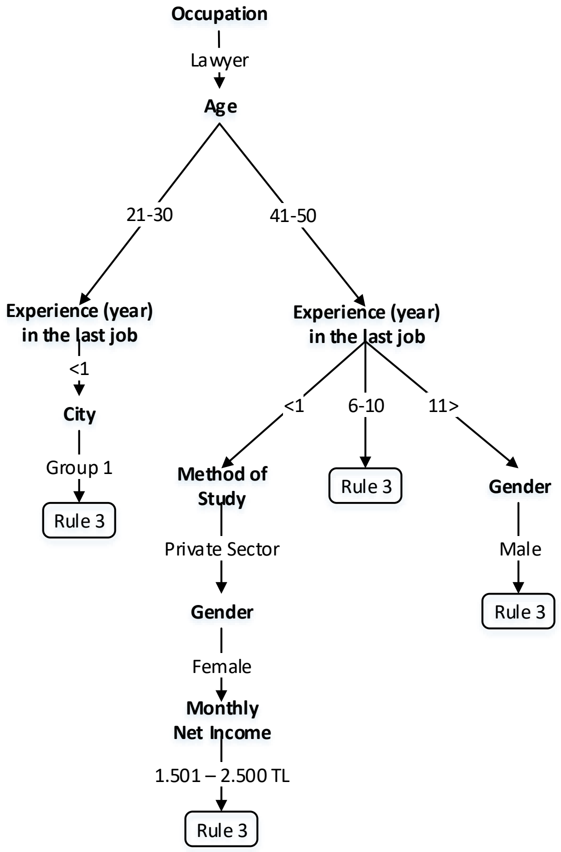

Fig 4. The decision tree structure of Rule 3

Rule 4: If (capital > 85.95) and (t-4 = D8) and (t-3 = S) and (t-2 = S) and (t-1 = S) = P (88, 5.93%)
Socio-demographic characteristics;

- (Occupation = 3) and (Age = 4) and (Experience (year) in the last job = 2) and (House that customers live= 3)
- (Occupation = 4) and (Gender = 2)
- (Occupation = 5) and (Age = 5) and (Educational Background= 7)

The decision tree structure of Rule 4 is given figure 5.

Rule 5: If (t-7 = S) and (t-6 = Y2) and (t-1 = D8) = P (71, 4.78%)
Socio-demographic characteristics;

- (Occupation = 3) and (Age = 3) and (Monthly Net Income = 2) and (Gender = 2)
- (Occupation = 3) and (Age = 3) and (Monthly Net Income = 3) and (Educational Background= 7)
- (Occupation = 5) and (Age = 4) and (Experience (year) in the last job = 2) and (Educational Background= 6) and (House that customers live= 1)

- (Occupation = 8) and (Educational Background= 2) and (Age = 2) and (Monthly Net Income = 4)

The decision tree structure of Rule 5 is given figure 6.

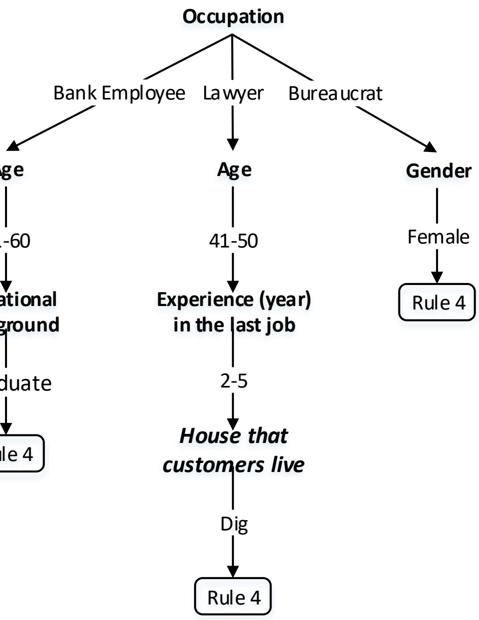

Fig 5. The decision tree structure of Rule 4

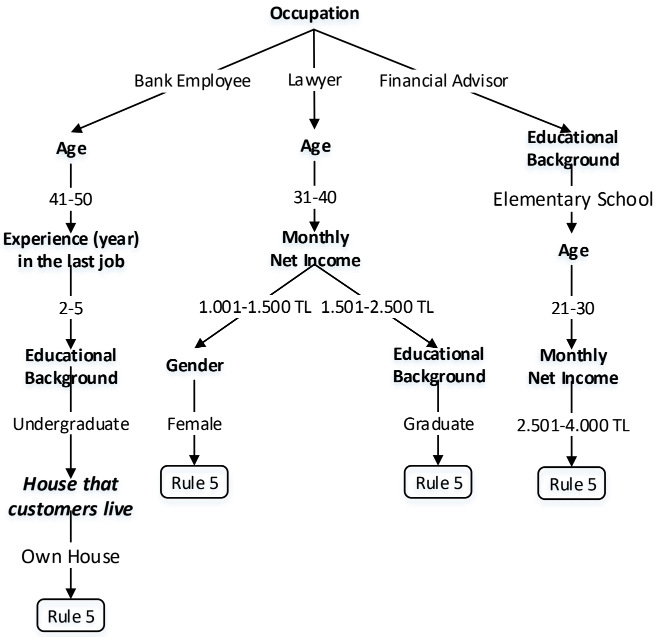

Fig 6. The decision tree structure of Rule 5

5. Conclusion

The most common investment account closure rule is rule 1. By following this rule, the percentage of customers that close their investment account is 33.13%. According to the rule, the customer who make transactions in line with D8 amount in t-2 month and make no transactions in t-1 month in other words the customer who sells at least 80% percentage of investment funds in t-2 month and do not make any transaction in t-1 month, will be closed his/her bank account in t month and will be evaded to be customer. If these people has no transaction in t-1 month and closed their investment accounts in month t, actually sold the entire investment funds in their accounts in month t-2. The customers behave fast to sell funds in their investment accounts and they only wait one month for closing their investment accounts. The bank does not have much time to contacting with these customers to convince them not to close the accounts. Thus, special attention is needed for this group. When we analysis socio-demographic data of these customers who close their investment accounts by applying this rule, we reach many different professions, different working types, and people of different age groups. One of the interesting features that a customer whose profession is 2 and 39 in other words soldiers and drivers is acting according to this rule.

The rule 2 is one of the most common investment account closure rule. By following this rule, the rate of customers who close their accounts is 23.84%. According to the rule the customer that not trading in t-7 and t-6 months, then trading in t-1 with D8 amount and that client closes it investment fund account. The customers, who close their investment accounts according to this rule, behave quickly than the customers that close their accounts in line with the rule 18. Similar to the rule 18, the bank does not have much time to contact with the clients to convince them to not to close their investment accounts. Special attention is required for this group. When we analysis socio-demographic data of these customers who close their investment accounts by applying this rule, we reach many different professions, different working types, and people of different age groups. The male customers whose profession is 4 in other words Minister, Deputy Bureaucrat acting according to this rule.

One of the most common investment rules to close an investment account is the rule 3 with 14.28% percentages. According to the rule, the customers operate in t-3 in line with D8 amount, not operating in t-2 and t-1 and in month, t closes the account and leaves being an investment fund client. This group waits one more month to close their accounts with respect to the rules 18 and 17, and this group also involves the customers who need special interest.

The customers who close their accounts according to the rule 4 act slowly than the customers that close the accounts according to the rules 1, 2, and 3. Moreover, the customers making transaction in month t-4 in line with D8 amount, and then wait in t-3, t-2, t-1 months with not making transactions and not closing their accounts. The customers whose professions are 17 in other words workers and the female customers whose profession is 4 in other words Minister, Deputy Bureaucrat acting according to this rule.

This work supports to identify the customers' socio-demographic characteristics, specialties of investments and the customers who turn to close the bank account can be determined. By this way, promotional activities would be lead to behave proactively in order not to lose the customers.

References

[1] Han, J., Kamber, M., Data mining: concepts and techniques 1st ed., Morgan Kaufmann, USA, 3-16, (2001).

[2] Chien C.-F., Chen, L.-F., Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry, Expert Systems with Applications, 34(1): 280-290 (2008).

[3] Giudici, P., Applied data mining: statistical methods for business and industry 1st ed., John Wiley & Sons, England, 1-15, 85-110, (2003).

[4] Fayyad, U., Piatetsky-Shapiro G., Smyth P., From data mining to knowledge discovery in databases, American Association for Artificial Intelligence, 3(17): 37-54 (1996).

[5] Apte, C., Weiss, S., Data mining with decision trees and decision rules, Future Generation Computer Systems, 13, 197–210 (1997).

[6] Fernandeza, I.B., Zanakisa, S. H., Walczakb, S., Knowledge discovery techniques for predicting country investment risk. Computers & Industrial Engineering, 43, 787–800. (2002).

[7] Witten, I., H., Frank, E., Data mining: practical machine learning tools and techniques 2nd ed., Morgan Kaufmann, USA, 62-415 (2005).

[8] Masand, B., Datta, P., Mani, D.R., Li, B., CHAMP: A prototype for automated cellular churn prediction, Data Mining and Knowledge Discovery 3, Netherlands, 219 - 225 (1999).

[9] Poel, D., and Lariviere, B., Customer attrition analysis for financial services using proportional hazard models, European Journal of Operational Research, 2004, vol. 157, No 1, 196-217.

[10] Ahn, J.H., Han, S.P., Lee, Y.S., Customer churn analysis: churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry, Science Direct, Telecommunications Policy, 30: 552 - 568 (2006).

[11] Ruta, D., Nauck, D., Azvine, B., K nearest sequence method and its application to churn prediction, Springer-Verlag, Berlin Heidelberg, 207 - 215 (2006).

[12] Chu, B.H., Tsai, M.S., Ho, C.S., Toward a hybrid data mining model for customer retention, Knowledge-Based Systems, 20 (8) (2007), pp. 703–718.

[13] Aydoğan, E., Gencer, C., Akbulut, S., Churn analysis and customer segmentation of a cosmetics brand using data mining techniques, Journal of Engineering and Natural Sciences, Sigma, 26 (2008).

[14] Gopal, R. K., Meher, S. K., Customer churn time prediction in mobile telecommunication industry using ordinal regression, Springer-Verlag, Berlin Heidelberg, 884 - 889 (2008).

[15] Farquad, M.A.H., Ravi, V., Raju, S.B., Data mining using rules extracted from svm: an application to churn prediction in bank credit cards, Springer-Verlag, Berlin Heidelberg, 390 - 397 (2009).

[16] Huang, B. Q., Kechadi, M. T., Buckley, B., Customer churn prediction for broadband internet services, Springer-Verlag, Berlin Heidelberg, 229 - 243 (2009).

[17] Chen, Z.-Y., Fan, Z.-P., Distributed customer behavior prediction using multiplex data: a collaborative MK-SVM approach, Knowledge-Based Systems, 35 (2012), pp. 111–119.

[18] Zhang, X., Zhu, J., Xu, S., Wan, Y., Predicting customer churn through interpersonal influence, Knowledge-Based Systems, 28 (2012), pp. 97–104.

[19] Slavescu, E., Panait, I., Improving customer churn models as one of customer relationship management business solutions for the telecommunication industry, “Ovidius” University Annals, Economic Science Series, 2012, vol. 12, Issue 1/201.

[20] Devi P., U., Madhavi, S., Prediction of churn behavior of bank customers using data mining tools, Business Intelligence Journal, 2012, vol.5, 96-101.

[21] Jamal, Z., Tang, H.K., Methods and systems for identifying customer status for developing customer retention and loyalty strategies, United States, Patent Application Publication, (2013).

[22] Verbeke, W., Martens, D., Baesens, B., Social network analysis for customer churn prediction, Applied Soft Computing, 2014, vol. 14, 431–446.

[23] Farquad, M.A.H., Ravi, V., Raju, S.B., Churn prediction using comprehensible support vector machine:An analytical CRM application, Applied Soft Computing, 2014, vol. 19, 31–40.

[24] Keramati, A., Jafari-Marandi, R., Aliannejadi, M., Ahmadian, I., Mozaffari, M., Abbasi, U., Improved churn prediction in telecommunication industry using datamining techniques, Applied Soft Computing, 2014, vol.24, 994–1012.