Modeling of user profiles of financial products and comparison of purchase prediction models: Based on machine learning

Rouyu Yu[1], Dongming Li[2,4] and Yining Liu[3]

[1]Oberlin College, Oberlin 44074, United States
[2]School of Science in Mathematics and Applied Mathematics, Xiamen University Malaysia, Negeri Selangor,43900, Malaysia
[3]Ulink College of Shanghai, Shanghai,201615, China

[4]MEC1909447@xmu.edu.my

Abstract. In this paperwork, the Apriori algorithm can be used to describe the association between data in a database. This paper will use the demographic and financial information of the users in the data set to analyze the user profile as the basic information. This work will fit the model by optimizing the Apriori algorithm, generating a combination of two or more associated variables, and then using the decision tree to learn and output the final model. The main contribution of this paper is to explore the optimization of Apriori and the effect of Apriori's combined model with the decision tree in predicting user buying behavior. Therefore, the conclusion can be drawn that in the sale of deposit products, if the customer has the above characteristics, such as no credit fault, no loan, housing loan, with a college degree, marital status is single, and act as an administrator. The people mentioned above may be more interested in the product. In other words, it is more likely to sell successfully. In future work, in addition to deposit products, more customer buying information is highly needed to extend predictable results to broader anger.

Keywords: Apriori, decision tree, association rule, methodology.

1. Introduction

In the past decade, Internet finance has developed rapidly, especially the customer's financial products. The financial industry is facing richer data and faster transactions. Given the technology nowadays, targeting specific customers with customized financial products is possible [1]. Banks can use statistical information to develop products or carry out marketing activities suitable for their target customers. Customized financial product advertising can aim directly at customers and provide unique personal and highly relevant experiences [1]. According to pretest results done by Walwave [2], in terms of advertising effectiveness, a moderate degree of personalization is expected to be the best. Personalized advertisements can increase click-through rates irrespective of whether banners appear on motive congruent or incongruent display websites [3], and people are almost twice as likely to click on an advertisement if it was customized. One of the downsides of personalized advertisements is the fact that it is more expensive than regular advertisements [4]. Traditional precision marketing in the

financial industry uses a large amount of manual labor to analyze data and acquire customers, but the customers of financial companies are divided into levels, and customers of the same level have similar characteristics.

Nowadays, the rise of machine learning has brought the possibility of lower customer acquisition costs and higher customer acquisition efficiency for financial companies. A recommendation system built using machine learning algorithms is a customer acquisition channel that financial companies can consider.A knowledge-based recommendation can be regarded as an inference technology to some extent [5]. It is not just recommended based on users' needs and preferences, but also uses rules for specific areas to carry out rule-based and case-based reasoning. For example, Burke [6] used the utility knowledge of restaurant dishes to recommend hotels to customers. Participatory new projects on the contours of financial information services are generally recommended based on strong recommendations based on common sense [7]. The recommendation algorithm based on common sense does not depend on the basics of customer scores, nor does it need to collect information about specific customers, because its identification has nothing to do with personal taste [8].

In the field of recommendation systems, association rules can analyze the user's habits and guide the system to make recommendations by mining and analyzing the association rules according to the user's profile. Based on correlation analysis and customer demographic information. According to the outcome of Apriori such as support and confidence results, it can be easily found that variables have a strong relationship with the successful or failed purchase rate, which will help us to make a good prediction on customer's buying behavior to give some personalized recommends.

In addition to the switching rules, the decision tree algorithm is also applied to establish the customer's personal information. Decision tree algorithm learning trainers, such as ID3 [9], build decision tree algorithms based on the recursive algorithm regions of the practice data information divided into subgroups until such subgroups only include individual cases [10]. In addition, the classification rules mined out by the tree are more precise to understand. Therefore,the decision tree has a fast computational speed as well as high accuracy. Kim et al. describes using tree induction techniques to match customer demographics with product categories to provide personalized describe the characteristics of training samples as "too accurate", it is impossible to realize a reasonable analysis of new samples, so it is not an optimal decision tree for analyzing new data [11]. Besides, the decision tree may be unstable, because a small change in the data may cause a completely different tree to be generated. This problem can be alleviated through the integration of decision trees. And if certain classes are dominant in the problem, the original decision tree will be biased, so it is recommended to balance the data set before fitting.

2. Data

This dataset is about the results of selling long-term deposits over the phone. It contains real data collected from a retail bank in Portugal between May 2008 and June 2013, with a total of 45,307 telephone contacts. The dataset is unbalanced because only 5091 successful sales have been completed. For a better assessment, this study divided the dataset into a training set (90%) and a test set (10%). The training set is used for model selection and training, including customers information performed as of June 2012, with a total of 41,187 examples.

The estimation of the decision tree algorithm is relatively small, and it is very easy to convert into a classification standard. The test set considers the predictive analytic capabilities of the selected entity model and contains 4,148 customer data from July 2012 to June 2013. Each record contains output overall goals and contact results ("No", "Yes"), alternative typing characteristics, telephone sales characteristics (such as call location), product details (such as the annual interest rate given), and customer information (For example, age, at work). This record is rich in the characteristics of the era and socioeconomic hazards (for example, the elasticity coefficient of the number of unemployed persons).

Table 1. Correlation

	y	age	dura-tion	cam-paign	m1_pdays	pre-vious	emp_var_rate	con_price_idx	cons_conf_idx	euri-bor3	nr_employed
y	1	0.01037	0.34951	-0.057	-0.324	0.20241	-0.2453	-0.12789	0.0447	-0.262	-0.2795
age	-0.01	1	-0.00	0.00747	-0.001	-0.00	0.04548	0.04468	0.11542	0.05482	0.04561
duration	0.34951	-0.00	1	-0.07	-0.080	0.03918	-0.0672	0.00272	0.00438	-0.075	-0.0932
cam-paign	-0.05	0.00747	-0.07	1	0.05412	-0.09	-0.1541	0.09463	-0.00214	0.13694	0.14201
m1_pdays	-0.32	-0.00	-0.08	0.05412	1	0.50729	0.22661	0.05849	0.08266	0.27757	0.28948
previous	0.20241	-0.00	0.03918	-0.091	-0.507	1	-0.4359	-0.28673	-0.11358	-0.454	-0.4393
emp_var_rate	-0.24	0.04548	-0.06	0.1541	0.22661	-0.43	1	0.66288	0.21997	0.9395	0.94622
cons_price_idx	-0.12	0.04468	0.00272	0.09463	0.05849	-0.28	0.66288	1	0.24276	0.48887	0.46748
cons_conf_idx	0.0447	0.11542	-0.00	-0.002	-0.082	-0.11	0.21997	0.24276	1	0.23313	0.12948
euri-bor3m	-0.26	0.05482	-0.07	0.13694	0.27757	-0.45	0.9395	0.48887	0.23313	1	0.92901
nr_employed	-0.27	0.04561	-0.09	0.14201	0.28948	-0.43	0.94622	0.46748	0.12948	0.92901	1
m1_job	0.04034	0.08005	0.00967	0.015	0.034	0.02616	-0.0529	-0.03334	0.02637	-0.051	-0.0464
m1_marital	0.03148	-0.21	0.00012	0.00302	-0.025	0.03518	-0.0490	-0.04153	-0.06119	-0.063	-0.0503
m1_education	0.04545	-0.08	-0.01	-0.000	-0.038	0.02238	-0.0154	-0.07348	0.0916	0.00356	-0.0040
m1_default	-0.10	0.19254	-0.01	0.04055	0.0792	-0.10	0.18094	0.16897	0.0376	0.17437	0.16112
m1_housing	0.01086	-0.00	-0.00	-0.009	-0.006	0.02513	-0.0529	-0.09629	-0.0706	-0.038	-0.0357
m1_loan	-0.00	0.00	-0.00	0.01805	-0.001	0.0033	0.00554	0.00083	0.0376	0.00738	0.0058
contact	0.14887	-0.03	0.03701	-0.068	-0.117	0.24153	-0.2272	-0.66282	-0.39809	-0.108	-0.1086
month	0.15815	0.06402	-0.01	-0.068	-0.117	0.11662	-0.0034	-0.2855	0.04802	0.05976	0.10914
day_of_week	0.00536	-0.02	-0.00	-0.011	0.00711	0.00711	-0.0132	0.00612	0.00156	0.00627	-0.0090
poutcome	0.2104	-0.00	0.04101	-0.092	0.53766	0.99746	-0.4369	-0.28878	-0.28878	-0.110	-0.1695

As can be seen from Table 1, the correlation coefficient of the variable 'outcome' and 'previous' is as high as 0.99, and by comparing the strength of their correlation with y, the variable of previous is chosen to be discarded. The correlation coefficients of 'emp_var_rate' and 'nr_employed' are 0.94,

and by comparing their strength and weakness with 'y' correlation, the variable 'emp_var_rate' is chosen to be discarded. The correlation coefficient between 'nr_employed' and euribor3m is 0.93, and by comparing the strength of their correlation with 'y', the variable 'euribor3m' is chosen to be discarded. At the same time, it can be concluded from the table, customer population information and property information and other distribution variables and 'y' correlation are not high, but at the same time these variables are needed to deal with these variables, one method can refer to the use of one-hot coding. Below a simple statistical analysis of age has proceeded, Duration and Job to understand the distribution of population information and find ways to deal with it.
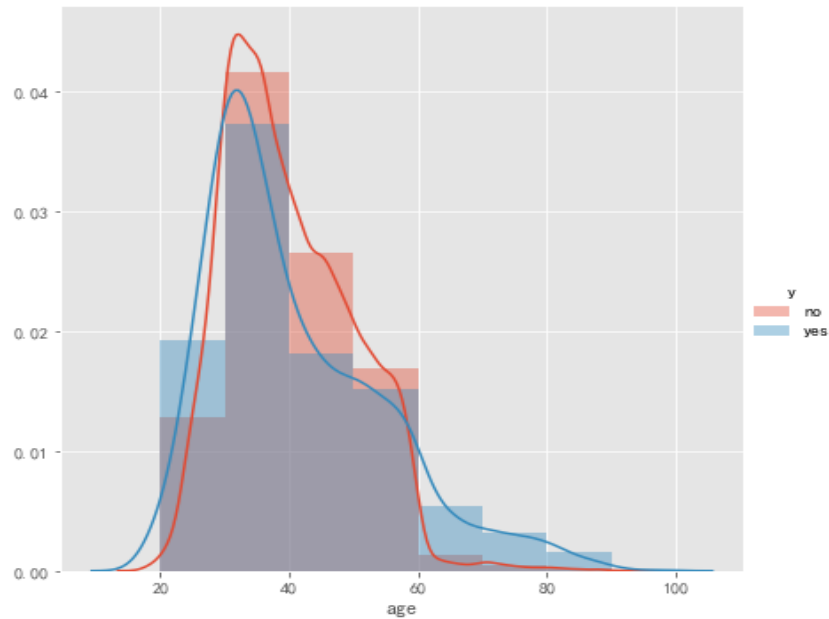

Figure 1. Age histogram.

As can be seen in Figure 1, only a small proportion of the bank's clients are under the age of 20 or below, with the age distribution concentrated in the middle-aged and young people aged 30 to 50, with an average age of 40.2 years. People in this age group have stable incomes, which may have an impact on the signing of products.
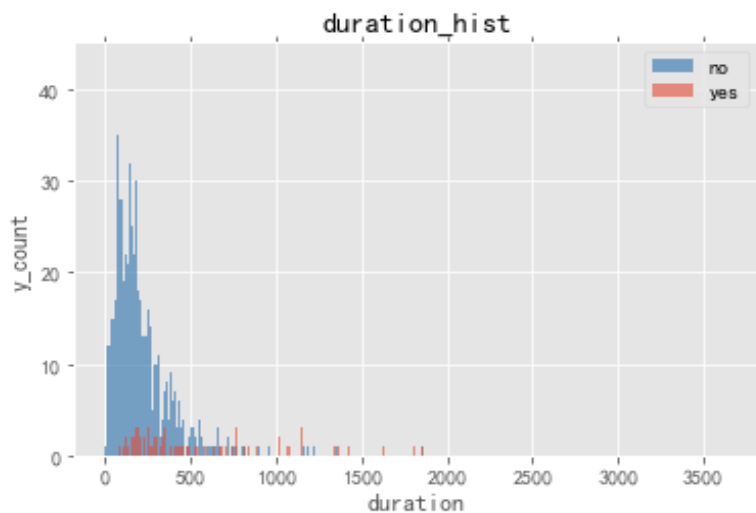

Figure 2. Duration distribution.

Customer call times who accept this deposit product are higher than that of the customer who did not accept, indicating that the variable may have a significant impact on the corresponding variable (Figure 2).
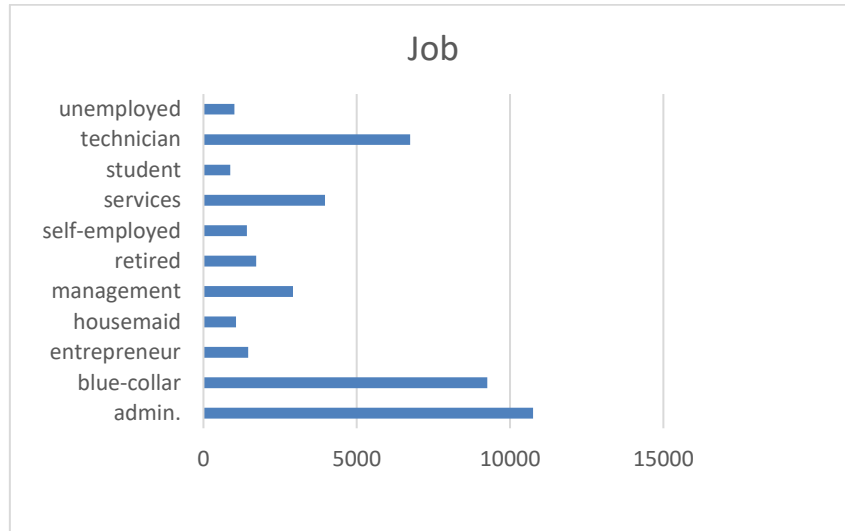

Figure 3. Job distribution.

Among the bank's customers. Executives, blue-collar workers, and technicians are in the top three, and the number of students, maids and unemployed are smaller, not to mention unknown jobs (Figure 3).

3. Methodology

The four-step methodology is used to determine purchase intent (Figure 4）

3.1. Step 1: correlation analysis

The correlation coefficient can reflect the relationship between two variables and their related directions. This can directly help us focus on some distinct highly related variables. However, the correlation coefficient cannot exactly demonstrate the relation degree between two variables. More analysis is still needed when making choices of variables.
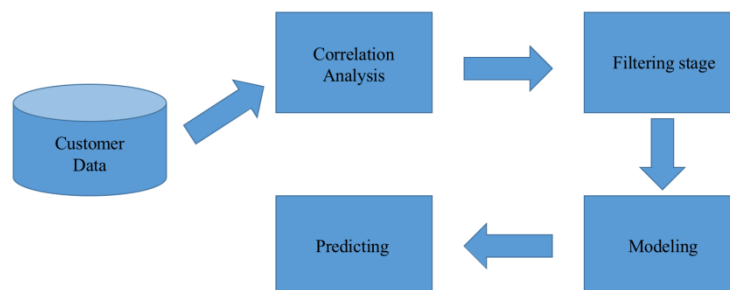

Figure 4. Methodology.

Notes: Dealing with the customer data, first of all, the correlation analysis is done to find some distinct highly related variables. Then in the filtering stage, the customer data is filtered and cleaned to conduct efficient computation. Next, in the modeling stage, the Apriori algorithm has done its job to describe the association between variables and the decision tree to learn and output the final model. Finally, in the prediction stage, customer purchase is made according to the user's intent based on the previous results.

3.2. Step 2: filtering stage

It is widely acknowledged that there exists useless information or abnormal values in the database. Because of that, filtering and cleaning the data is necessary for efficient computation.

3.2.1. Identification of outliers. Usually, the identification of outliers can be done using graphical methods (such as box plots, normal distribution diagrams) and modeling methods (such as linear regression, clustering algorithm, K nearest neighbor algorithm). In this research, two Graphical methods and a method of identifying outliers are shared based on models.

3.2.2. Box plot. The box plot technique is used to section off quantiles of data to identify outliers. It's a typical statistical graph and is widely used in academia and industry. The lower quantile in the figure refers to the value corresponding to the 25% quantile of the data ($Q1$); the median is the value corresponding to the 50% quantile of the data ($Q2$); the upper quartile is the number corresponding to the 75% quantile of the data ($Q3$); the calculation formula for the upper whisker is $Q3 + 1.5\,(Q3 - Q1)$; the calculation formula for the lower whisker is $Q1 - 1.5\,(Q3 - Q1)$. Among them, $Q3 - Q1$ represents the interquartile range. If a box plot is used to identify outliers, the criterion for judging is when the data value of the variable is greater than the upper whisker of the box plot or smaller than the lower whisker of the box plot, such a data point can be considered an abnormal point. Therefore, based on the box plot above, the abnormal points and extreme abnormal points in a numerical variable can be defined, and their judgment expressions are shown in table 2:

Table 2. (Judgment expression)

Criteria	Result
$x > Q3 + 1.5(Q3 - Ql)\ or\ x < Ql - 1.5(Q3 - Ql)$	Outlier
$x > Q3 + (Q3 - Ql)\ or\ x < Ql - 3(Q3 - Ql)$	Extreme Outlier

Using the matplotlib module in Python to visualize data is vital, where the boxplot function is used to draw box plots. Taking the data of the number of sunspots from 1700 to 1988 as an example, the box plot method is used to identify outliers in the data (Figure 5)
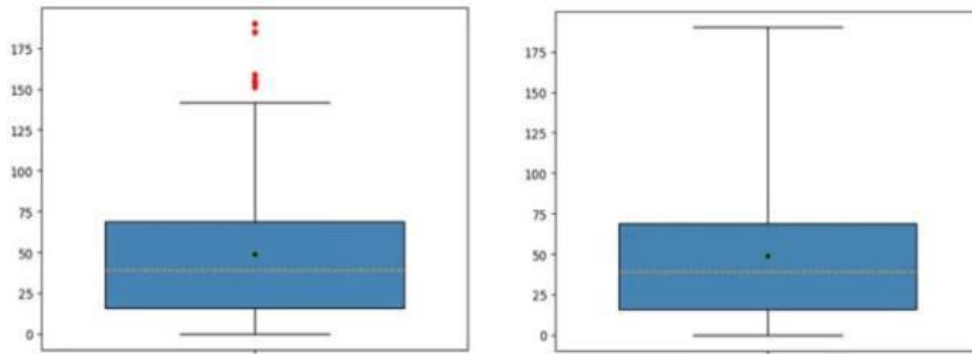

Figure 5. Example of the box plot.

Note: As shown in the figure above, the boxplot function in the matplotlib submodule plot can be used to draw box plots very conveniently. The top and bottom of the left figure must be set to 1.5 times the interquartile range, and the top and bottom of the right figure must be set to 3 times The interquartile range. It can be seen from the left image that there are at least 5 abnormal points in the data set, all of which are above the upper whiskers; however, no extreme abnormal points are shown in the right image.

3.2.3. Normal distribution diagram. According to the definition of standard normal distribution, the probability that data information points fall within 1 standard deviation (sigma value) of the average positive and negative poles is 68.2%; data information points fall within 2 relative standard deviations

of the average positive and negative poles. The probability is 95.4%; the probability that the data information point falls within 3 standard deviations from the mean positive and negative is 99.6%.

Therefore, consider the probability value mentioned above from another perspective. If the probability that a data information point falls outside the two standard deviations of the average positive and negative poles is less than 5%, it is accidental, that is, that information point is considered an abnormal point. Similarly, if the data information point falls outside the three standard deviations of the average positive and negative poles, the probability will be even smaller. This kind of data information point can be regarded as an abnormal point due to extremes. In order to better let the reader, understand the probability value mentioned in the article content, the probability density graph of the standard normal distribution can be inquired, as shown in the figure below (Figure 6):
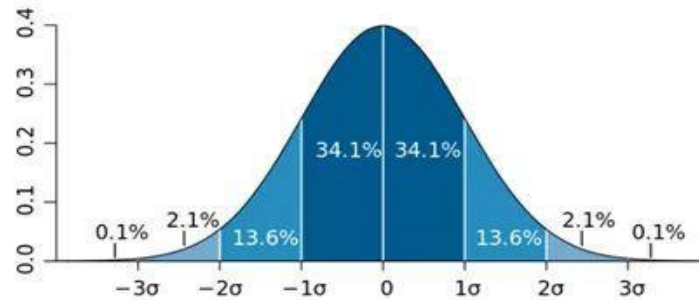

Figure 6. Probability Density diagram of the standard normal distribution.

Further, based on the conclusion of the above figure, the abnormal points and extreme abnormal points of the numerical variables can be identified according to the judgment conditions in the following table (Table 3):

Table 3. Judgment conditions.

Criteria	Result
$x > !x + 2\sigma$ or $x < !x - 2\sigma$	Outlier
$x > !x + 3\sigma$ or $x < !x - 3\sigma$	Extreme Outlier

Using the knowledge points of normal distribution, combined with the plot function in the Pyplot sub-module, draw line graphs and scatter plots, and identify outliers or extreme outliers with the help of two horizontal reference lines.

3.3. Step 3: modeling

Apriori algorithm is used to describe the association between variables, fitting the model by optimizing the Apriori algorithm, generating a combination of two or more associated variables, and then using the decision tree to learn and output the final model [12].

3.3.1. Apriori. Apriori [13] is an optimization algorithm for the discovery of current itemsets and association rules learning and training on relational databases. It is based on identifying a new item that recurs in the database query and expands them to a large new item set, as long as this new item set occurs repeatedly in the database query. As everyone knows, the Apriori optimization algorithm has two shortcomings. The first is that the number of itemsets converted into alternatives is likely to be very large. The second is the time-intensive daily task of measuring applicability. Scanner item set database query over and over [14].

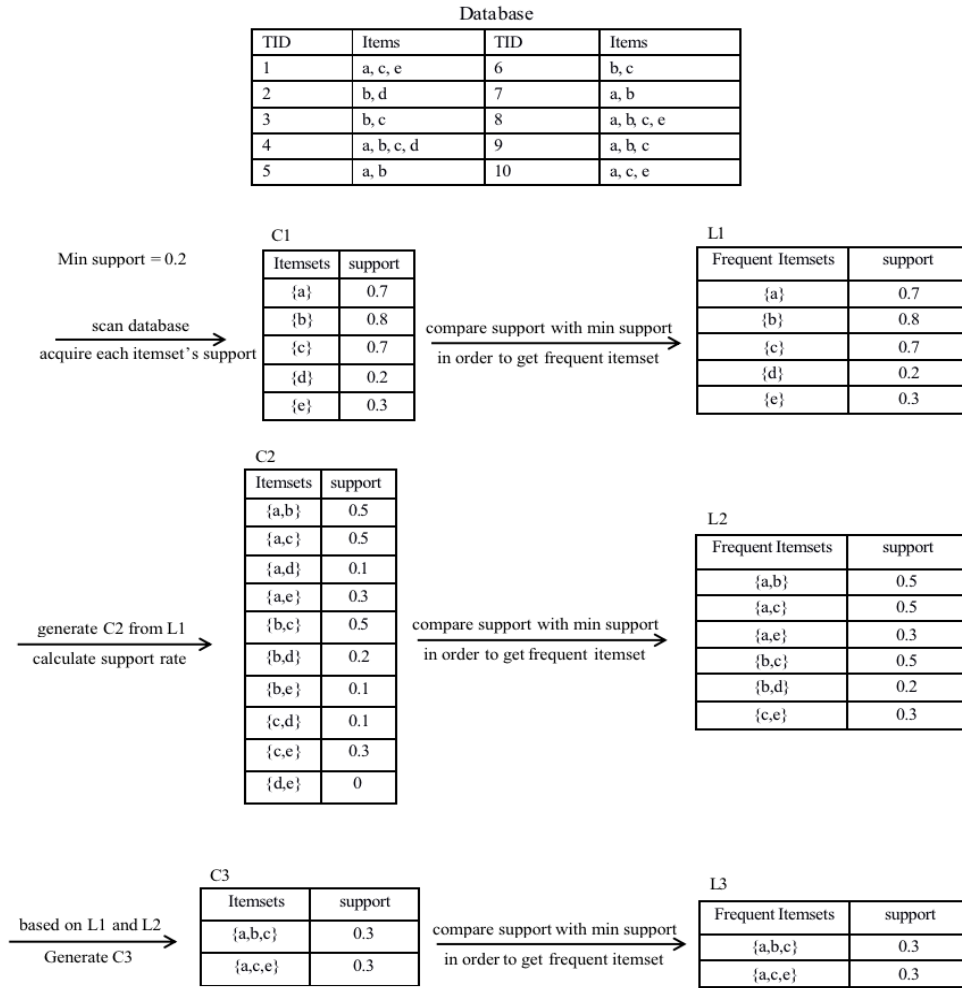Figure 7 is an example of Apriori shown below:

Database

TID	Items	TID	Items
1	a, c, e	6	b, c
2	b, d	7	a, b
3	b, c	8	a, b, c, e
4	a, b, c, d	9	a, b, c
5	a, b	10	a, c, e

Min support = 0.2

C1

Itemsets	support
{a}	0.7
{b}	0.8
{c}	0.7
{d}	0.2
{e}	0.3

scan database
acquire each itemset's support

compare support with min support
in order to get frequent itemset

L1

Frequent Itemsets	support
{a}	0.7
{b}	0.8
{c}	0.7
{d}	0.2
{e}	0.3

C2

Itemsets	support
{a,b}	0.5
{a,c}	0.5
{a,d}	0.1
{a,e}	0.3
{b,c}	0.5
{b,d}	0.2
{b,e}	0.1
{c,d}	0.1
{c,e}	0.3
{d,e}	0

generate C2 from L1
calculate support rate

compare support with min support
in order to get frequent itemset

L2

Frequent Itemsets	support
{a,b}	0.5
{a,c}	0.5
{a,e}	0.3
{b,c}	0.5
{b,d}	0.2
{c,e}	0.3

C3

Itemsets	support
{a,b,c}	0.3
{a,c,e}	0.3

based on L1 and L2
Generate C3

compare support with min support
in order to get frequent itemset

L3

Frequent Itemsets	support
{a,b,c}	0.3
{a,c,e}	0.3

Figure 7. Example Apriori.

Note: Example of apriori algorithm flowchart.

Association rule

Association rule analysis is a technique to uncover how items are associated with each other. There are three common ways to measure association [15].

Support

This says how popular an itemset is, as measured by the proportion of transactions in which an itemset appears.

$$Support\ \{X\} \ = \ \frac{\{X\}}{number(all\ samples)} \qquad (1)$$

Confidence

This says how likely item Y is purchased when item X is purchased, expressed as $\{X -> Y\}$. This is measured by the proportion of transactions with item X, in which item Y also appears.

$$Confidence\{X \rightarrow Y\} \ = \ \frac{Support\{X,Y\}}{Support\{X\}} \qquad (2)$$

Lift up

This shows the probability of buying product Y when buying product X and at the same time manipulates product Y's popularity level. A thrust value exceeding 1 indicates that if item X is purchased, it is likely to choose shopping item Y, while a value below 1 indicates that if item X is purchased, it is unlikely to choose shopping item Y.

$$Lift\{X \rightarrow Y\} = \frac{Support\{X, Y\}}{Support\{X\} \times Support\{Y\}} \tag{3}$$

Main steps

Step 1: Calculate the compatibility of the object. The combination in the query of the transaction management database ($size\ k = 1$) (note that the suitability is the frequency of the occurrence of the item set). This is called transforming into a candidate set.

Step 2: Cut the alternative set according to the project whose applicability of removal is lower than the given threshold (Figure 8)

Step 3: Add a regular itemset to generate a combination of $size\ k$ 1, and repeat the above combination until no more itemsets can be generated. This situation occurs when the resulting combination has lower applicability than the given applicability [15].
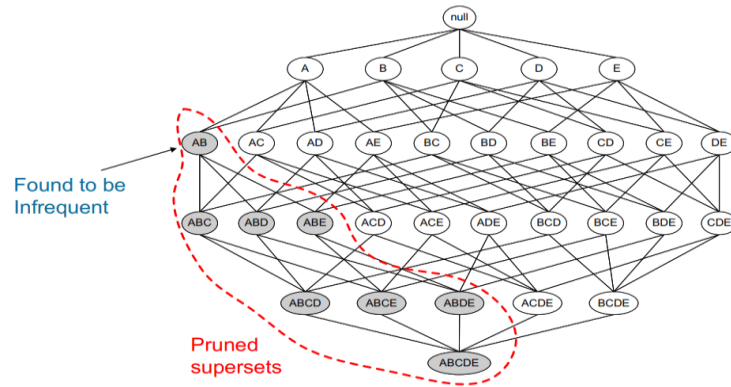

Figure 8. Apriori Pruning [10].

Note: It started from every single item in the item setlist. Then, the candidates are generated by self-joining. Extending the length of the item sets one item at a time. The subset test is performed at each stage and the item sets that contain infrequent subsets are pruned. Repeating the process until no more successful item sets can be derived from the data is very important.

3.3.2. Decision tree algorithm. The decision tree algorithm is based on the known occurrence probability of various situations. By constructing the decision tree, the probability that the expected value of the net present value is greater than or equal to zero is obtained, the project risk is evaluated, and its feasibility is judged. This is a decision analysis method that intuitively uses probability analysis. There are two main disadvantages of the decision tree algorithm: first, it is prone to overfitting. Second, the correlation of attributes in the data set is easily ignored.

After a period of development trend, various decision tree algorithms have been improved, their characteristics have been gradually improved, and they can solve various types of data information. Some major optimization algorithms will be discussed below.

ID3

The ID3 (Iterative Dichotomiser 3) algorithm is developed by Quinlan [9]. It uses information gained to select the appropriate attributes for each node of the generated decision tree. The attribute with the highest information gain (the greatest degree of entropy reduction) is used as the test attribute of the current node. In such a way, the information required to classify the subset of training samples obtained from the later division will be minimal.

CART

The classification and regression tree (CART) is proposed by Breiman et al. [14]. This is a non-parametric decision tree learning technology that generates either a classification tree or regression tree based on whether the dependent variable is categorical or numerical. It uses the Gini index for choosing attributes.

C4.5

The C4.5 algorithm is developed by Quinlan [9]. It is an improved version of the ID3 algorithm. This algorithm uses information gain as a splitting standard. It can handle missing attribute values, which is an improvement from the ID3 algorithm

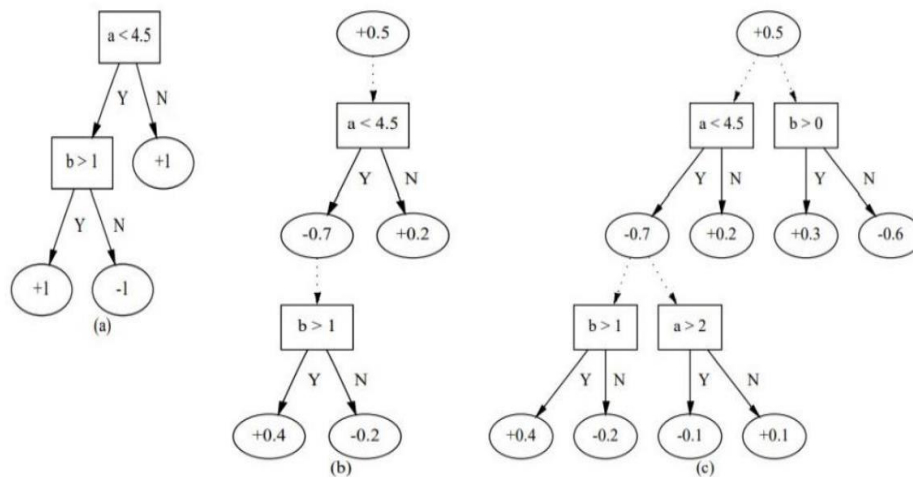Figure 9 is an example of tree algorithm shown below:

Figure 9. Example of a tree algorithm [14].

3.4. Step 4: predicting

Based on Apriori results, it can be concluded that people with certain characteristics are more likely to buy or not to buy the product. Then, using these variables to make customers' profiles predict their purchase intent is highly recommended. Next, putting the variables from Apriori into Decision Tree algorithm to process and forecast models which can also help predict customers' purchase intent [14].

4. Result

In Apriori analysis results (Table 4,5), all the lift values' are greater than 1, which means variables are related tight with each other. Then, sorted the confidence values in descending order and pick the leading terms. It is obvious that customers who are willing to purchase the deposit products usually have the following characteristic: have no credit in fault, do not have a loan, have a housing loan, have a university degree, marital status is single, and work as an admin. On the contrary, customers who are not willing to purchase the products usually have married marital status, have no credit in fault, have no housing loan, work as blue-collar or technician, and have basic 4 or 9 education or high school degree [14].

Table 4. Apriori results (1).

Antecedents	Consequents	Support	Confi-	Lift
frozen-	frozenset(('default_no'})	0.112654	1	1.000072
frozen-	frozenset({'loan_no'})	0.096071	0.852801	1.005300
frozen-	frozenset({*housing_yes'})	0.063465	0.563362	1.028261
fro2enset({*y_	frozen-	0.053535	0.475215	1.037044
frozen-	frozenset(('education_university.degree'})	0.046639	0.414008	1.226864
frozen-	frozen-	0.039720	0.352586	1.238366

Table 4. (continued).

frozen-set({'y yes'})	frozenset({'marital single'})	0.03933 1844	0.34913 7931	1.243 10970 8
frozen-set(('y yes*})	fro2enset({*job_admin.'})	0.03372 3415	0.29935 3448	1.146 74198 5
frozen-set({*y_yes*})	frozen-set({*housing_yes', 'default_no\ 'education_university.degree'})	0.02614 8393	0.23211 2069	1.242 23387 4
frozen-set({*y_yes*})	frozen-set(('marital_married*, 'education_university.degree'})	0.02318 6365	0.20581 8966	1.132 11425 6
frozen-set({*y_yes*})	frozen-set({'marital_married\ 'default_no', *education_university.degree'})	0.02318 6365	0.20581 8966	1.132 11425 6
fro2enset({*y_yes*})	frozen-set(('housing_yes\ 'marital_single', *default_no*})	0.02260 3671	0.20064 6552	1.277 11793 7
frozen-set({*y_yes*})	frozen-set({'loan_no', 'housing_yes\ 'education_university.degree'})	0.02214 2372	0.19655 1724	1.266 71450 7
frozen-set({'y_yes'})	frozenset({'loan_no', 'housing_yes\ default_no', 'education_university.degree'})	0.02214 2372	0.19655 1724	1.266 71450 7
frozen-set({'y_yes'})	frozen-set({'job_admin.\ *default_no*, 'education_university.degree'})	0.02158 3956	0.19159 4828	1.277 3402
frozen-set({'y_yes*})	frozen-set({'housing_no*, 'education_university.degree'})	0.02049 1405	0.18189 6552	1.207 79544 9
frozen-set({'y_yes'})	frozenset({*housing_no', default_no' 'education_university.degree'})	0.02049 1405	0.18189 6552	1.207 79544 9

Table 5. Apriori Results (2).

Antecedents	Consequents	Support	Confidence	Lift
frozen-set((·y no·})	frozcnsct((·marital marricd·})	0.545401 573	0.614643 756	1.012313 941
fro-zensct({·y no·})	frozcn-set((·marital marricd\ 'loan no\ *dcfault no'})	0.462537 632	0.521259 713	1.011764 612
frozen-set({·y no·})	frozcnset({'housing no*})	0.402932 893	0.454087 775	1.004347 937
frozen-set({·y no·})	froz-cnsct({^aritaLmarricd*, 'housing ycs\ 'loan^o*, *dcfault no'})	0.246503 836	0.277799 059	1.008460 042
frozen-set((·y no·})	frozcnset({'job bluc-collaF })	0.209187 142	0.235744 774	1.049260 401
fro-zensct({·y no·})	frozcnsct({*cduca tion high. school*})	0.205982 325	0.232133 085	1.004844 721
frozen-set({·y no·})	frozcnsct({'job tcchnician'})	0.145989 123	0.164523 367	1.004951 568
frozen-set({·y no·})	frozcnsct({·loan yes·})	0.135112 169	0.152265 514	1.003763 122
frozen-set((·y no·})	frozcnsct({*loan ycs\ 'dcfault no'})	0.135112 169	0.152265 514	1.003763 122
fro-zensct({·y no·})	frozcn-set((·marital marricd·, 'loan no\ *dcfault no', 'job b luc-collar'})	0.129163 834	0.145562 001	1.053304 231
frozen-set({·y no·})	frozcn-set({*loan no\ 'job technician\ 'dcfault no'})	0.124575 119	0.140390 719	1.005287 367
frozen-set({·y no·})	frozcnsct({'loan no·，'cducation basic.9y'})	0.115664 757	0.130349 13	1.040065 859
frozen-set((·y no·})	froz-cnsct({*loan no\ *dcfault no', 'cducation basic.9y'})	0.115664 757	0.130349 13	1.040065 859

Table 5. (continued).

fro-zensct({·y no· })	froz-cnsct((·marital marricd\ 'cducation high.schoor })	0.114013 79	0.128488 563	1.023237 999
frozen-set({·y no·})	frozcnsct({'housing ycs' 'cducation high.school'}) ,	0.111245 994	0.125369 377	1.002273 663
frozen-set({·y no·})	froz-cnsct({·marital marricd·, *dcfault no', *cducation basic.9y'})	0.093813 732	0.105723 979	1.045763 512
frozen-set((·y no·})	frozcnsct({'loan noL 'housing^ycs*, cducation high.school'}) ,	0.093182 48	0.105012 586	1.004705 784
fro-zensct((·y_no' })	frozcnsct({*cducation_basic.4y'})	0.090997 378	0.102550 071	1.011454 102

Then, users can use the variables from the results to create customers' profiles (Figure 10) and put those characteristics into cloud words in order to make the results more obvious.

Figure 10. Cloud words of customer characters [16].

Note: The left cloud words are customers who are willing to purchase the deposit product and the right cloud words are those not willing to purchase [16].

5. Conclusion

Using apriori algorithm to create customer profiles, it can be found that customers who are willing to purchase the deposit products usually have the following characteristics: have no credit in fault, do not have a loan, have a housing loan, have a university degree, marital status is single, and work as an admin. Therefore, it can be easily concluded that in the deposit products selling, if the customers own the above characteristics, they may be more interested in the products. In other words, it will be more likely to sell the products successfully. In future work, more customer purchase information is highly needed besides the deposit products so as to extend our predictable results to a wider range.

References

[1] Walrave, M., Poels, K., Antheunis, M.L., Van den Broeck, E., van Noort, G. (2018) Like or dislike? Adolescents' responses to personalized social network site advertising. Journal of Marketing Communications. 24: 599-616.

[2] Semerádová, T., Weinlich, P. (2019) Computer estimation of customer similarity with Facebook lookalikes: Advantages and disadvantages of hyper-targeting. IEEE Access. 7: 153365-153377.

[3] Bleier, A., Eisenbeiss, M. (2015) Personalized online advertising effectiveness: The interplay of what, when, and where. Marketing Science. 34: 669-688.

[4] Marquis, A. (n.d.) Definition of transparency advertising. https://smallbusiness.chron.com/definition-transparency-advertising-35939.html.

[5] Felfernig, A., Jeran, M., Ninaus, G., Reinfrank, F., Reiterer, S. (2013) Toward the next generation of recommender systems: Applications and research challenges. In: Tsihrintzis, G.A., Virvou, M., Jain, L.C. (Eds.), Multimedia services in intelligent environments: Advances in recommender systems. Springer International Publishing, Heidelberg, pp. 81-98.

[6] Burke, R. (2000) Knowledge-based recommender systems. Encyclopedia of Library and Information Systems. 69: 180–200.

[7] Chung, R., Sundaram, D., Srinivasan, A. (2007) Integrated personal recommender systems. In: Proceedings of the ninth international conference on Electronic commerce. Minneapolis, MN, USA. pp. 65–74.

[8] Quinlan, J.R. (1986) Induction of decision trees. Machine Learning. 1: 81-106.

[9] Pazzani, M.J., Billsus, D. (2007) Content-based recommendation systems. In: Brusilovsky, P., Kobsa, A., Nejdl, W. (Eds.), The adaptive web: Methods and strategies of web personalization. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 325-341.

[10] Kim, J.W., Lee, B.H., Shaw, M.J., Chang, H.-L., Nelson, M. (2001) Application of decision-tree induction techniques to personalized advertisements on internet storefronts. International Journal of Electronic Commerce. 5: 45-62.

[11] Agrawal, Rakesh, et al.(1998) Fast Algorithms for Mining Association Rules, Fast Algorithms for Mining Association Rules and Readings in Database Systems 3rd Ed.,97-102.

[12] Moro, S., Cortez, P., Rita, P. (2014) A data-driven approach to predict the success of bank telemarketing. Decision Support Systems. 62: 22-31.

[13] Agrawal, R., Srikant, R., (1994). Fast algorithms for mining association rules. Proceedings of the 20th International Conference on Very Large Data Bases, Santiago, Chile, pp. 487-499.

[14] Breiman, L., Friedman, J., Olshen, R., Stone, C. (2017) Classification and regression trees.

[15] Chonyy. (2020) Apriori: Association rule mining in-depth. https://towardsdatascience.com/apriori-association-rule-mining-explanation-and-python-implementation-290b42afdfc6.

[16] Freund, Y., Mason, L. (1999) The alternating decision tree learning algorithm. In: International Conference on Machine Learning. pp. 124-133.