Customer Profiling Method with Big Data based on BDT and Clustering for Sales Prediction

Kohei Arai[1], Zhan Ming Ming[2], Ikuya Fujikawa[3],Yusuke Nakagawa[4], Tatsuya Momozaki[5], Sayuri Ogawa[6]
Information Science Dept, Saga University, Saga City, Japan[1]
SIC Co., Ltd, Hakata-ku, Fukuoka City, Fukuoka, Japan[2, 3, 4, 5, 6]

***Abstract*—We propose a method for customer profiling based on Binary Decision Tree: BDT and k-means clustering with customer related big data for sales prediction; valuable customer findings as well as customer relation improvements. Through the customer related big data, not only sales prediction but also categorization of customers as well as Corporate Social Responsibility (CSR) can be done. This paper describes a method for these purposes. Examples of the analyzed data relating to the sales prediction, valuable customer findings and customer relation improvements are shown here. It is found that the proposed method allows sales prediction, valuable customer findings with some acceptable errors.**

Keywords*—Customer profiling; binary decision tree: BDT; corporate social responsibility (CSR); k-means clustering; sales prediction; valuable customer findings

I. INTRODUCTION

When "analyzing customer data", it must be understood what kind of data exists, firstly. In addition, since the data in the database is used for various purposes, it is not always stored in a format suitable for analyzing customer data. Therefore, the first thing to do is to get an overall picture of customer data by profiling. Here, we propose to assume a certain data format and think based on that data format.

This is a data retention format that is generally used as target data for data mining. The retention format is a simple table with customer numbers in the vertical rows and all possible customer attributes and indicators in the horizontal columns. This table may be physically created as a database table, or it may be considered as an image when starting the analysis work just by assuming such a format.

Since the target of the analysis is the customer, first, this format is used so that the information about each customer is listed in each line. On the other hand, the attributes and index values arranged in the column are called variables. It is a "variable" meaning a "changing numerical value" because it takes a different value for each customer. The variable means the standard for looking at the customer and defines "from what point of view the customer is understood".

To perform profiling, let's look at each variable independently. Each variable is classified as either a quantitative variable or a qualitative variable. In the case of quantitative variables, understand the characteristics of variable values by using representative values such as average value, minimum value, and maximum value. Also, by understanding the distribution, it becomes possible to understand how the variable values are concentrated / distributed.

Once we understand the characteristics of each variable by profiling a single variable, the next step is to understand the relationships between the variables. A scatter-like visualization of each customer in dots makes it possible to understand whether the two variables are in direct proportion, inversely proportional, or completely random. When deciphering the relationship between variables, it is necessary to proceed with the analysis with a certain purpose, not just a visual and intuitive grasp. Data mining can be considered as one of the methods.

Large analytical datasets are vertically compressed by age group. In other words, customers are classified based on a single variable (here, age), and the tendency of each classification (= segment) is grasped. In addition to this, we have added a row for the population and a new index for the number of customers. This allows us to compare with the whole customer and understand the size of each segment. At this time, since each row to be analyzed is aggregated from customer (single customer) to segment (multiple customers), the variables to be used are average value, minimum value, maximum value, total value, composition ratio, etc. It will be converted into an index and compared. And the difference in the index value is the characteristic of the segment.

If it can be understood what kind of products it is contracting / purchasing for each segment, what kind of channel it is using, what kind of usage pattern it is using, etc. It can be understood whether it has such characteristics. With this, the profiling work is almost completed, and we have grasped what kind of variable characteristics each customer group has. The next task is analysis for "behavior". Considering that the target of analysis is the customer, the "behavior" here is an action to the customer, specifically a campaign activity.

Although there are many methods for customer profiling, performance is not good enough. Also, it is rare the method which allows prediction of sales based on the customer profiling. In this paper, we intend to improve the customer profiling performance and to propose a method for sales prediction based on Deep Learning.

There are prediction related research works as follows,

Probabilistic cellular automata-based approach for prediction of hot mudflow disaster area and volume is proposed [1]. Also, new approach of prediction of Sidoarjo hot mudflow disaster area based on probabilistic cellular automata is proposed [2]. These prediction methods are expanded to GIS based 2D cellular automata approach for prediction of forest fire spreading [3].

Cell based GIS as Cellular Automata: CA for disaster spreading prediction and required data systems is proposed [4]. The method is applied to hot mudflow prediction area model and simulation based cellular automata for LUSI and plume at Sidoarjo East Jawa [5].

Comparative study between Eigen space and real space-based image prediction methods by means of autoregressive model is conducted [6]. Also, comparative study on image prediction methods between the proposed morphing utilized method and Kalman filtering method is conducted [7].

Another prediction method for time series of imagery data in Eigen space is proposed [8] together with image prediction method with non-linear control lines derived from Kriging method with extracted feature points based on morphing [9]. Cell based GIS as cellular automata for disaster spreading predictions and required data systems are developed [10].

Prediction method of El Nino Southern Oscillation: ENSO event by means of wavelet-based data compression with appropriate support length of base function is proposed [11]. On the other hand, Question Answering: Q/A for collaborative learning with answer quality prediction is created [12].

Wildlife damage estimated and prediction using blog and tweet information is conducted in [13]. Meanwhile, prediction method for large diatom appearance with meteorological data and MODIS derived turbidity as well as chlorophyll-a in Ariake Bay area in Japan is proposed [14].

Method for thermal pain level prediction with eye motion using SVM is proposed in [15]. Meanwhile, prediction method for large diatom appearance with meteorological data and MODIS derived turbidity and chlorophyll-a in Ariake bay area in Japan [16].

Smartphone image based agricultural product quality and harvest amount prediction method is proposed [17]. On the other hand, data retrieval method based on physical meaning and its application for prediction of linear precipitation zone with remote sensing satellite data and open data is proposed and validated with the actual data [18].

Recursive Least Square: RLS method-based time series data prediction for many missing data is proposed [19]. Meanwhile, prediction of isoflavone content in beans with Sentinel-2 optical sensor data by means of regressive analysis is proposed and validated with the actual data [20].

In this connection, we propose a method for customer profiling based on Binary Decision Tree: BDT and K-means clustering with customer related big data for sales prediction and valuable customer findings as well as customer relation improvements. The scikit-learn of BDT[1] is used in the study. K-means clustering of scikit-learn[2] is also used.

The following section describes research background followed by the proposed method. Then example of the experimental actual data of sales is described as a validation of the proposed method followed by conclusion with some discussions.

II. Research Background

A. Importance of Customer Profiling

Customer profiling is important for improving sales environment and customer relation. In order to create customer profiles, customer clustering is needed, first. Then valuable customers could be found. These activities help customer need surveillance together with customer satisfaction as shown in Fig. 1.

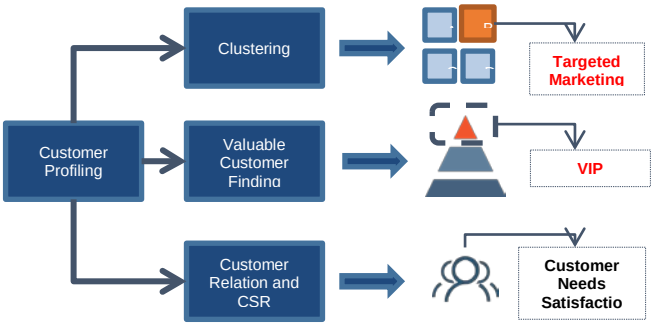


Fig. 1. Importance of Customer Profiling.

B. Parameters for Customer Profiling

There are 7 layers of the parameters for customer profiling as shown in Table I. In the layers, typical data labels as shown in Fig. 2 must be considered.

The parameters in each layer are shown in Fig. 2, the parameters in Fig. 2 shows all the possible parameters, some of them are influencing to sales prediction. The first layer includes customer information while the second layer includes demographics information. Meanwhile, the third layer includes geographic information while the fourth layer includes information on customers' preference and interest. On the other hand, the fifth layer includes shopping pattern while the sixth layer includes brand affinity. The seventh layer includes purchase action related information such as risk to lose, propensity to buy, predicted date to come.

TABLE I. Parameters for Sales Prediction

Tier 7. Prediction	**7th**
Tier 6. Brand Affinity	**6th**
Tier 5. Shopping Pattern	**5th**
Tier 4. Preference & Interests	**4th**
Tier 3. Geographic	**3rd**
Tier 2. Demographics	**2nd**
Tier 1. Recognition & Contact	**1st**

[1] https://scikit-learn.org/stable/modules/tree.html
[2] https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1

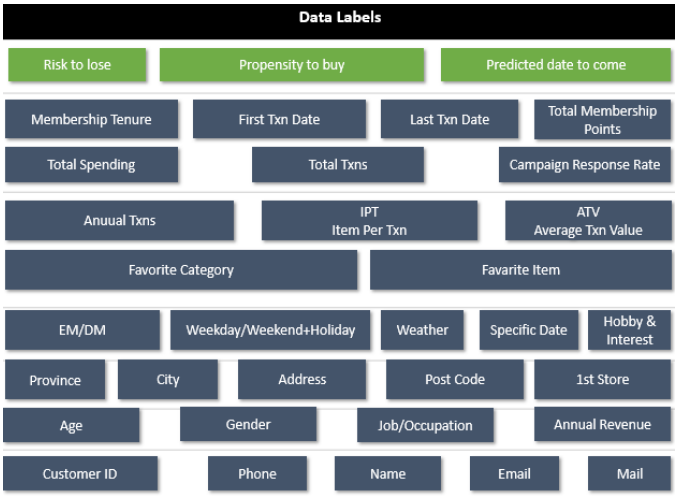


Fig. 2. Data Labels (Parameters) in the Assigned Layers.

III. Proposed Sales Prediction Method

A. Sales Prediction

Fig. 3 shows the proposed process flow of sales prediction. The customer profile, including the information on gender, age, address, occupation, education etc. must be made available first. Then it is followed by segmentation. After that, AI models are created, and finally total sales is going to be predicted.

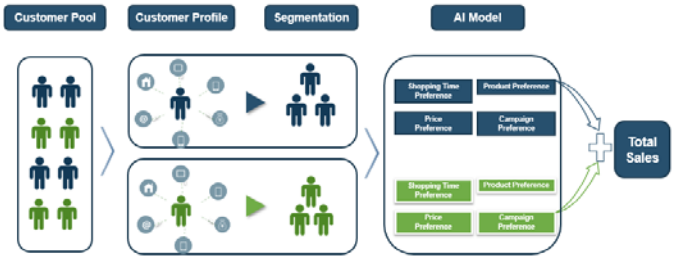


Fig. 3. Proposed Process Flow of Sales Prediction.

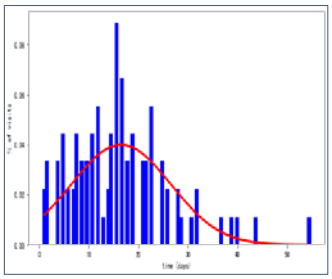
Fig. 4 also shows the details of sales prediction targets. The next visit date, types of sales products and sales are predicted with the AI models. Time, product, price preferences are modeled by AI through learning processes of deep learning.

Fig. 4. Details of Sales Prediction Targets.

B. Sales Prediction with Deep Learning

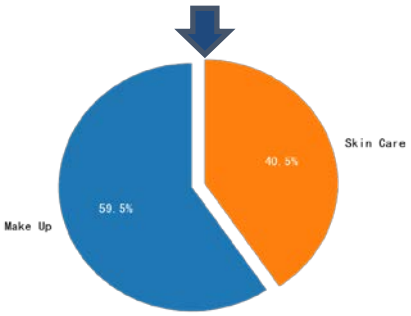
Fig. 5 shows the preferences of the time, the product, and the price for the specific customer. The customer is specified through the previous processes with deep learning. All the preferences are obtained from the sales information such as shown in Table II. This is a just example of the specific customer's sales data. Total sales can also be predicted through the learned deep learning. In the proposed Deep Learning, TensorFlow Keras is used. Two hidden layers are implemented with the number of neurons of 250. Also, ReLU is used for activation function together with the fully connected layer with SoftMax function.

cust_id	tran_date	Date Gap
18647	20180719	22
18647	20180804	16
18647	20190823	384
18647	20190920	28
18647	20191011	21
....	....	



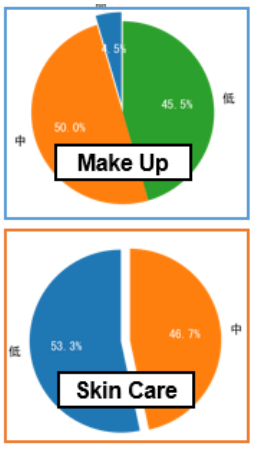
(a)Time preference

Product Name	No. of Sales	% Ratio
Make Up	22	59.5%
Skin Care	15	40.5%



(b)Product Preference.

ProductName	ProductPrice	%pct
Make Up	High	4.5%
Make Up	Middle	45.5%
Make Up	Low	50.0%
Skin Care	Middle	46.7%
Skin Care	Low	53.3%



(c)Price Preference.

Fig. 5. Preferences of the Time, the Product and the Price.

IV. Experimental Results from Sales Prediction

A. Preparation of Data for Deep Learning

The first thing we must do is to check the missing data in the customer information. Fig. 6 shows the percentage ratio of the missing data by item by item. The customer rank is the most frequently missing data item followed by e-mail addresses of the customer. Therefore, we must conduct deep learning processes without the missing data. The second thing we must do is to standardization of sales data. The ranges are different from each other so that the sales data are standardized with mean and standard deviation so as to adjust the mean and standard deviation are 0 and 1 after the standardization.

Fig. 7 shows the percentage ratio of the missing data in the sales data. As shown in Fig. 7, the highest percentage ratio of the missing data about sales is motivation to visit the shop followed by the staff ID who made services to the customer. These missing data of customer related information and the sales data must be care about in the learning processes.

TABLE II. Sales Data for the Specific Customer

Customer ID	Product ID	Product Price	No. of Sales
133034	32	Skin Care	1
133034	48	Skin Care	1
133034	80	Skin Care	1
133034	25	Make Up	1
133034	33	Make Up	1
133034	42	Make Up	1
133034	79	Make Up	2
133034	82	Make Up	3
133034	1	Make Up	2
133034	4	Make Up	2
133034	7	Make Up	1
133034	12	Skin Care	1
133034	19	Make Up	1
133034	24	Skin Care	1
133034	29	Make Up	1
133034	36	Make Up	1
133034	37	Skin Care	1
133034	40	Make Up	2
133034	52	Skin Care	1
133034	54	Make Up	1

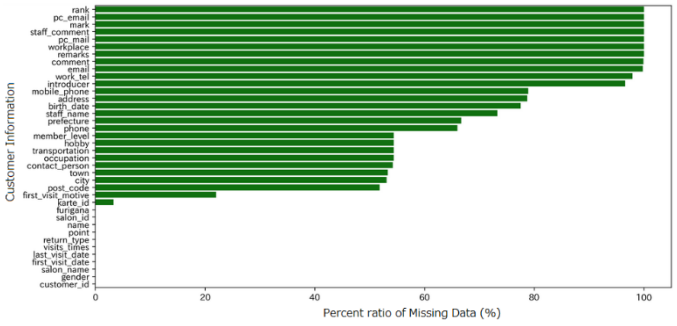


Fig. 6. Percentage Ratio of the Missing Data in the Sales Data.

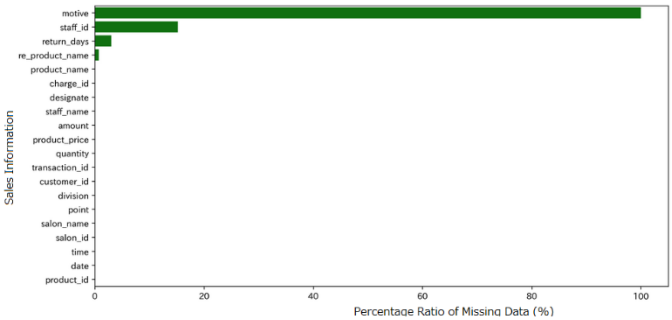


Fig. 7. Percentage Ratio of the Missing Data in the Sales Data.

Fig. 8 shows rating results of the percentage ratio of the sales type (Product name). The highest sales type is the adult hair cut (sales type No.1) followed by the gray hair dyeing (sales type No.2).

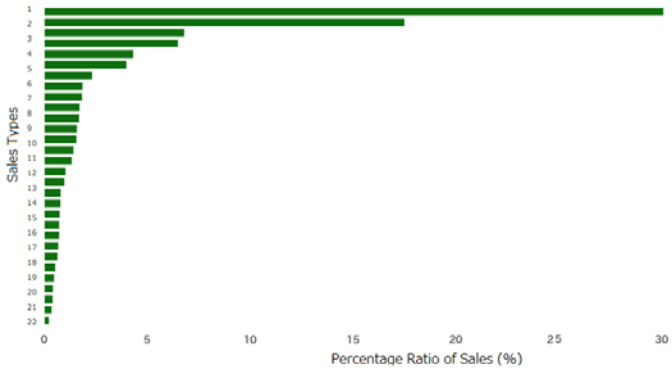


Fig. 8. Rating Results of the Percentage Ratio of the Sales Type (Product Name).

Monthly sales can also be calculated from the sales data. Fig. 9 shows the monthly sales for the specific hair salon for the period from September of 2010 to December of 2021.

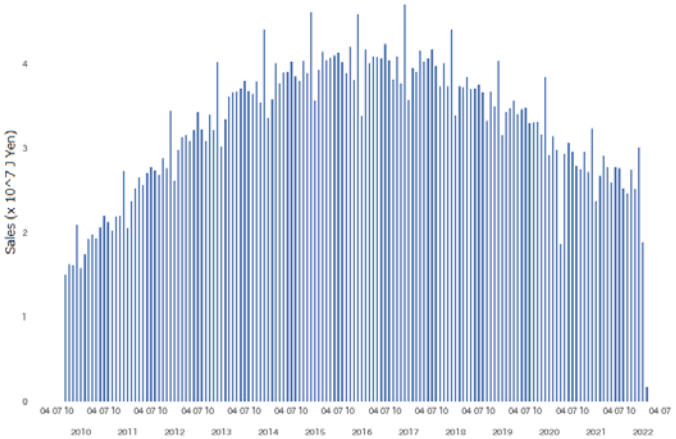


Fig. 9. Monthly Sales for the Specific Hair Salon.

Fig. 10 shows the revisit day interval (repeat cycle) for the specific hair salon.

As shown in Fig. 10, highest frequency ranges from 30 to 60 days. Therefore, it is found that the customer visits the specific hair salon every month to two months. Also, histogram of customers' age distribution is shown in Fig. 11. There are

two peaks at around 15 to 20 years old for both male and female customers and 45 to 85 years old for female customers. This is the second layer of the customer profile, demographic information. The other layer data are also calculated from the sales data.

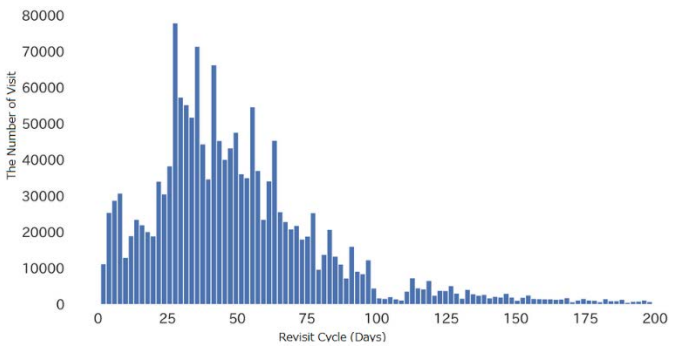


Fig. 10. Revisit Day Interval (Repeat Cycle) for the Specific Hair Salon.

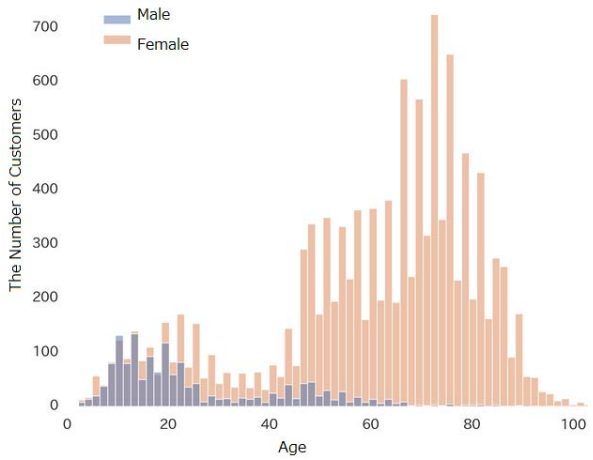


Fig. 11. Histogram of Customers' Age Distribution.

B. Customer Clustering

The number of visits to the specific hair salon is investigated as one of the parameters (feature vector) of the customer clustering.

Fig. 12 shows relation between the number of visit and the number of customers (green bars) as well as percentage ratio of revisit customers (red solid line). More than 95% of customers are repeat customers as shown in Fig. 12.

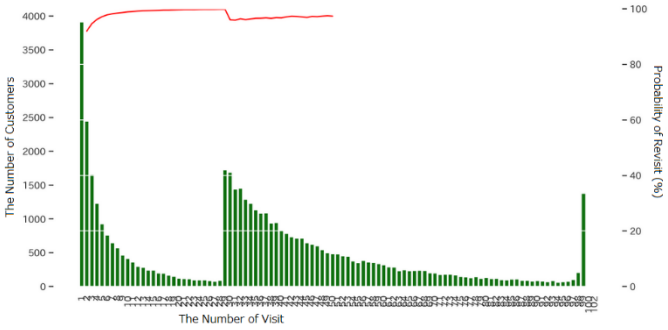
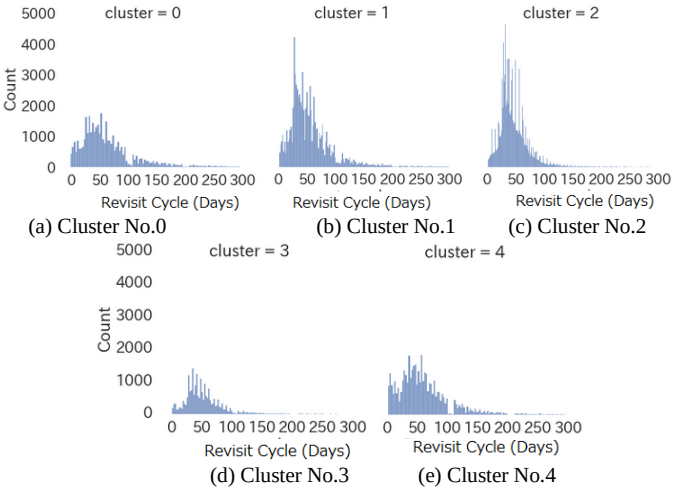


Fig. 12. Relation between the Number of Visit and the Number of Customers as well as Percentage Ratio of Revisit Customers.



(a) Cluster No.0 (b) Cluster No.1 (c) Cluster No.2

(d) Cluster No.3 (e) Cluster No.4

Fig. 13. Histogram of the Revisit Cycle (Days).

Then K-mean clustering is applied to the sales data with the number of clusters is five. Histogram of the revisit cycle (days) is shown in Fig. 13. This is one of the results from the clustering. Also, the learning processes are done with deep learning for each cluster. The cluster No. 0 is resembling to the cluster No. 4.

TABLE III. NUMBER OF CUSTOMERS FOR EACH CLUSTER

Cluster No.	The_Number_of_Customers
1	1833
4	1439
2	1360
0	1316
3	564

There is the peak of the histogram at the revisit cycle around 50. On the other hand, the cluster No. 1 is like the cluster No. 2 the peaks of the histograms of the cluster No. 1 and 2 are much higher than that of the cluster No.0 and 4 as well as the cluster No.3. The number of customers for each cluster is shown in Table III.

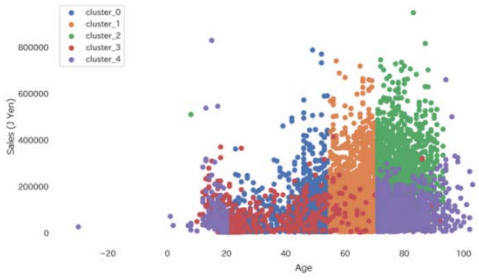
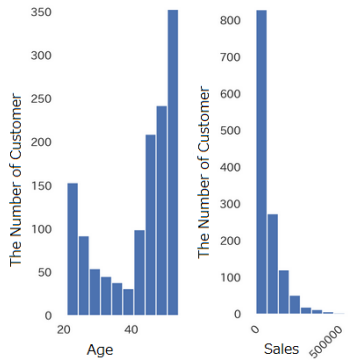


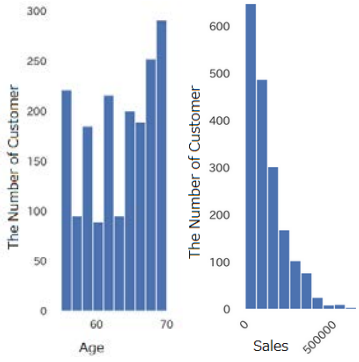
Fig. 14. Relation between the Sales Amount and Age.

C. Clustering Results

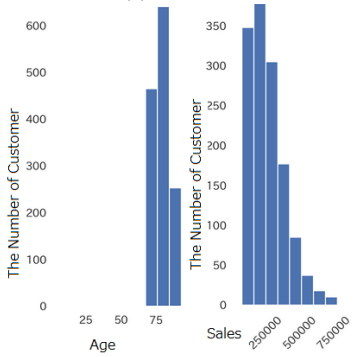
One of the clustering results is shown in Fig. 14 (relation between the age and the sales amount) for each cluster. Also, Fig. 15 shows the relations between the number of customer and age as well as sales for each cluster. As shown in Fig. 14 and 15, all the customers are divided into five clusters clearly and these clusters are well characterized with their profiles.



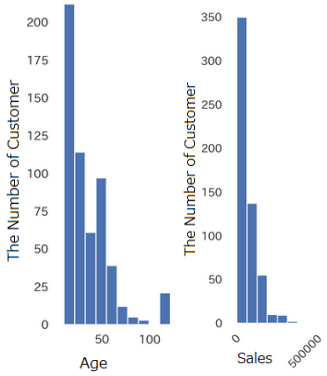
(a) Cluster 0



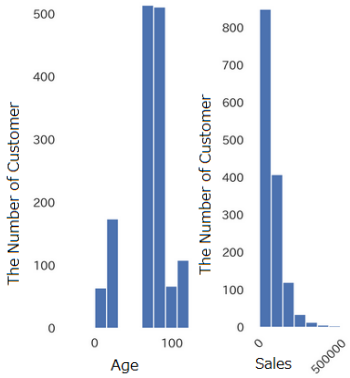
(b) Cluster 1



(c) Cluster 2



(d) Cluster 3



(e) Cluster 4

Fig. 15. Relations between the Number of Customer and Age as well as Sales.

Fig. 16 shows the sales amount for each cluster. The sales amount of the cluster No. 2 is highest followed by cluster No. 1. Also, there is the big dip at the late of 2019 since the number of customers is getting down by the influence due to the COVID-19.

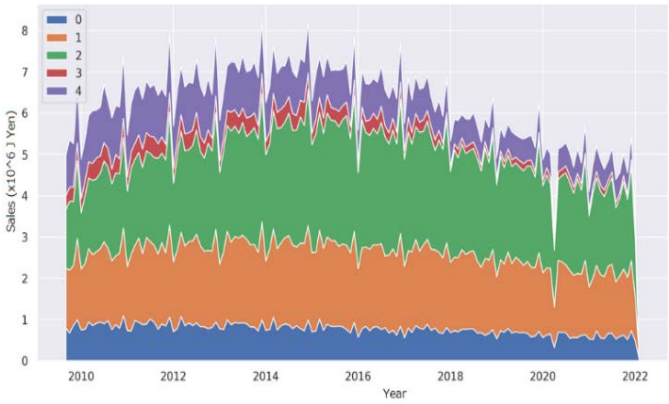


Fig. 16. Shows the Sales Amount for each Cluster.

D. Deep Learning

After the clustering, all the customers are divided into five clusters. Then the customers are segmented. Also, the number of transactions (visit), the sales amount is investigated. TensorFlow of deep learning is used for sales prediction. The training sample data is the sales data of the year from 2010 to 2013 and the validation data is the sales data of the year of 2014. The four years learning processes are conducted and then validation is done for the year of 2014. The validation result is shown in Fig. 17.

As shown in Fig. 17, prediction accuracy is not good enough (Root Mean Square Error: RMSE= 17.04: 21.25%) since the training sales data is not enough for deep learning. Also, customers' behavior is changed by year so that the different customers' behavior between 2010 to 2014 and 2014 induces such prediction error.

Fig. 17. Sales Prediction Result for the Year of 2014 with the Learning Processes during 2010 and 2013.

V. Conclusion

We proposed a method for customer profiling based on Binary Decision Tree: BDT and K-means clustering with customer related big data for sales prediction and valuable customer findings as well as customer relation improvements. Through the customer related big data, not only sales prediction but also categorization of customers as well as Corporate Social Responsibility: CSR can be done.

This paper describes a method for these purposes. Examples of the analyzed data relating to the sales prediction, valuable customer findings and customer relation improvements are shown in this paper. It is found that the proposed method allows sales prediction, valuable customer findings with some acceptable errors (21.25% of RMSE).

VI. Future Reaesrch Work

Further investigations are required for improvement of prediction accuracy.

Acknowledgment

The authors would like to thank to Professor Dr. Hiroshi Okumura and Professor Dr. Osamu Fukuda for their valuable discussions.

References

[1] Achmad Basuki, Tri Harsono and Kohei Arai, Probabilistic cellular automata based approach for prediction of hot mudflow disaster area and volume, Journal of EMITTER1, 1, 11-20, 2010.

[2] Kohei Arai, Achmad Basuki, New Approach of Prediction of Sidoarjo Hot Mudflow Disaster Area Based on Probabilistic Cellular Automata, Geoinformatica - An International Journal (GIIJ), 1, 1, 1-11, 2011.

[3] Kohei Arai, Achmad Basuki, GIS based 2D cellular automata approach for prediction of forest fire spreading, International Journal of Research and Reviews on Computer Science, 2, 6, 1305-1312, 2011.

[4] Kohei Arai, Cell based GIS as Cellular Automata for disaster spreading prediction and required data systems, CODATA Data Science Journal, 137-141, 2012.

[5] Kohei Arai, A.Basuki, T.Harsono, Hot mudflow prediction area model and simulation based cellular automata for LUSI and plume at Sidoarjo East Jawa, Journal of Computational Science (Elsevior) 3,3, 150-158, 2012.

[6] Kohei Arai, Comparative Study between Eigen Space and Real Space Based Image Prediction Methods by Means of Autoregressive Model, International Journal of Research and Reviews in Computer Science (IJRRCS) Vol. 3, No. 6, 1869-1874, December 2012, ISSN: 2079-2557.

[7] Kohei Arai, Comparative Study on Image Prediction Methods between the Proposed Morphing Utilized Method and Kalman Filtering Method, International Journal of Research and Reviews in Computer Science (IJRRCS) Vol. 3, No. 6, 1875-1880, December 2012, ISSN: 2079-2557.

[8] Kohei Arai Prediction method for time series of imagery data in eigen space, International Journal of Advanced Research in Artificial Intelligence, 2, 1, 12-19, (2013).

[9] Kohei Arai Image prediction method with non-linear control lines derived from Kriging method with extracted feature points based on morphing, International Journal of Advanced Research in Artificial Intelligence, 2, 1, 20-24, (2013).

[10] Kohei Arai, Cell based GIS as cellular automata for disaster spreading predictions and required data systems, Advanced Publication, Data Science Journal, Vol.12, WDS 154-158, 2013.

[11] Kohei Arai, Prediction method of El Nino Southern Oscillation event by means of wavelet based data compression with appropriate support length of base function, International Journal of Advanced Research in Artificial Intelligence, 2, 8, 16-20, 2013.

[12] Kohei Arai, Anik Nur Handayani, Question Answering for collaborative learning with answer quality prediction, International Journal of Modern Education and Computer Science, 5, 5, 12-17, 2013.

[13] Kohei Arai, Shohei Fujise, Wildlife Damage Estimated and Prediction Using Blog and Tweet Information, International Journal of Advanced Research on Artificial Intelligence, 5, 4, 15-21, 2016.

[14] Kohei Arai, Prediction method for large diatom appearance with meteorological data and MODIS derived turbidity as well as chlorophyll-a in Ariake Bay area in Japan, International Journal of Advanced Computer Science and Applications IJACSA, 8, 3, 39-44, 2017.

[15] Kohei Arai, Method for Thermal Pain Level Prediction with Eye Motion using SVM, International Journal of Advanced Computer Science and Applications IJACSA, 9, 4, 170-175, 2018.

[16] Kohei Arai, Prediction method for large diatom appearance with meteorological data and MODIS derived turbidity and chlorophyll-a in Ariake bay area in Japan, International Journal of Advanced Computer Science and Applications IJACSA, 10, 9, 39-44, 2019.

[17] Kohei Arai, Osamu Shigetomi, Yuko Miura, Satoshi Yatsuda, Smartphone image based agricultural product quality and harvest amount prediction method, International Journal of Advanced Computer Science and Applications IJACSA, 10, 9, 24-29, 2019.

[18] Kohei Arai, Data Retrieval Method based on Physical Meaning and its Application for Prediction of Linear Precipitation Zone with Remote Sensing Satellite Data and Open Data, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 11, No. 10, 56-65, 2020.

[19] Kohei Arai, Kaname Seto, Recursive Least Square: RLS Method-Based Time Series Data Prediction for Many Missing Data, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 11, No. 11, 66-72, 2020.

[20] Kohei Arai, Prediction of Isoflavone Content in beans with Sentinel-2 Optical Sensor Data by Means of Regressive Analysis, Proceedings of SAI Intelligent Systems Conference, IntelliSys 2021: Intelligent Systems and Applications pp 856-865, 2021.

Authors' Profile

Kohei Arai, He received BS, MS and PhD degrees in 1972, 1974 and 1982, respectively. He was with The Institute for Industrial Science and Technology of the University of Tokyo from April 1974 to December 1978 also was with National Space Development Agency of Japan from January, 1979 to March, 1990. During from 1985 to 1987, he was with Canada Centre for Remote Sensing as a Post Doctoral Fellow of National Science and Engineering Research Council of Canada. He moved to Saga University as a Professor in Department of Information Science on April 1990. He was a councilor for the Aeronautics and Space related to the Technology Committee of the Ministry of Science and Technology during from 1998 to 2000. He was a councilor of Saga University for 2002 and 2003. He also was an executive councilor for the Remote Sensing Society of Japan for 2003 to 2005. He is a Science Council of Japan Special Member since 2012. He is an Adjunct Professor of University of Arizona, USA since 1998. He also is Vice Chairman of the Science Commission "A" of ICSU/COSPAR since 2008 then he is now award committee member of ICSU/COSPAR. He wrote 77 books and published 670 journal papers as well as 500 conference papers. He received 66 of awards including ICSU/COSPAR Vikram Sarabhai Medal in 2016, and Science award of Ministry of Mister of Education of Japan in 2015. He is now Editor-in-Chief of IJACSA and IJISA. http://teagis.ip.is.saga-u.ac.jp/index.html