A Two-Stage Customer Journey Analytical Model in Single House Business

Sotarat Thammaboosadee[1], **Benjathip Chinomi**[2], and **Ehab K. A. Mohamed** [3]

ABSTRACT

The single housing industry is currently experiencing a continuous expansion in demand for housing. Addressing the needs of different customer groups is the key to increasing the sales rate. The objective of this research is to propose a single house customer journey analytical model that consists of two stages. The first stage concerns the customer journey between registration and reservation processes. The second one identifies the customer loyalty from the reservation to the transfer stage. We experimented with four classification data mining techniques. The experimental results include comparison of the accuracy and F-Measure. We also performed statistical testing. The Artificial Neural Network with 1 hidden layer presented the highest accuracy and F-measure of 0.970 and 0.958 in the stage of reservation and 0.882 and 0.862 in the stage of transference, compared to other algorithms. This model analyzes the probability of the customer progressing through the stages to the conclusion of purchase by learning the customer's characteristics and the factors involved in the customer's decision. The model displays the reservation and transfer result for customers who have achieved the respective reservation and transference steps according to their registration profile. Experiments showed that the proposed two-stage models could predict customer loyalty, thereby enhancing relationship management between customers and organizations. It also confers a competitive advantage within the industry.

Keywords: Real Estate, Customer Relationship Management, Customer Journey, Data Mining, Single House

1. INTRODUCTION

The situation of the residential market is revealed in statistical data and survey results of demand and supply of the residential market. There is an expansion of demand in the residential market caused by a reduction in the transfer and mortgage fees [1]. The information in the residential market shows that the business is highly competitive. Therefore, getting in touch with customers is the key to gaining a competitive advantage. Real estate firms are engaged in the sale of various types of houses and land for sale. There is development of residential projects in flat landscape areas in the form of single houses, semi-detached houses, and townhouses. The construction occurs together with the development of public utilities in the project. The variety of choices exists not only to cover all levels of prices and to meet the needs of different customer groups, but also to enhance the quality of life of customers and to deliver good things to society through the development.

A residence is one of the four necessities a human needs. By the theory of four requisites, humans need a residence to protect their bodies from winds, rains, and dangers. In the past, humans built their permanent house located close to the river in order to use the water for planting, drinking, and bathing, which differs from the present [2]. The selection of a residence is crucial, and there are several factors to consider before paying for it. Factors affecting the development model of the project include rising land prices, selling prices, customer age, design and function of the product, and also the customer's income.

Therefore, classifying the customers into groups is an advantage in order to understand and track them on their journey [3]. Data mining [4] can be an advantage in this issue by classifying data about customer characteristics and several factors which will affect the decision of customers.

The objective of this research is to predict customer loyalty along the customer journey in the real estate business, specific to single house products. The scope of this two-stage customer journey consists of registration-to-reservation and reservation-to-transfer stages using classification models to predict customer loyalty. The model is used to determine a focus group of customers by predicting the probability of a customer buying the product. This research can help the real estate business to establish customer loyalty and create a competitive advantage.

Manuscript received on March 25, 2020 ; revised on April 30, 2020.

Final manuscript received on May 01, 2020.

[1]The author is withFaculty of Engineering, Mahidol University, Thailand. E-mail: sotarat.tha@mahidol.ac.th

[2]The author is with Celestica Inc., Thailand. E-mail: sunnychinomi@gmail.com

[3]The author is with Faculty of Management Technology, German University in Cairo Cairo, Egypt. E-mail: ehab.kamel@guc.edu.eg

DOI: 10.37936/ecti-cit.2020142.240239

2. RELATED THEORIES

2.1 Real Estate Business Process

Real estate property, in legal terms, means land or other assets that are attached to the land, such as townhouse condominiums, commercial buildings and dormitories. The definition in business refers to activities and services that are sold under the prevailing legal statutes [5]. The characteristics of real estate can be divided in three major categories.

1) General trading

General trading is the stock that is traded similarly to other businesses, such as buying and then selling. The law requires transactions in writing and registered with the official documents such as property deeds showing the ownership. That is because real estate is a highly-valued asset. It must be controlled in order to avoid legal problems.

2) Rent

Examples of rent include renting an apartment, rooms, houses, and warehouses. When investors buy or build the buildings, they also buy land and then lease it out. It will generate a fixed income.

3) Brokerage housing or rental

This category is the most familiar. However, each trade broker may earn up to 100 percent, depending on the capabilities and other factors.

According to the high sales volume, the demand of customers has increased, and in the process of selling the house is divided into several parts. Each process will vary depending on the completeness of the home sold. The main processes consist of reserve, loan, and transfer. In order to boost sales in the property business, sales transactions will be divided into two parts: sales and transfers. Unfortunately, customers can make changes at every step of the process, such as booking, and cancelling if customers have no ability to get a loan. Once the customer has cancelled, the business loses the sales opportunity that leads to the final stage of the sale.

2.2 Customer Relationship Management

Customer Relationship Management (CRM) [6] is a business strategy to build relationships with clients, learn customer needs, and respond to customer needs with a product or service that meets the key customer relationship management needs. CRM enables organizations to improve customer relationships, increase revenue, reduce costs, supply customers, and enhance customer satisfaction by modeling and product development to meet customer needs. Increasing the quality of customer relationship management is critical to managing and determining customer satisfaction. Standards in companies such as customer data collection, channel management, and product and service development will be identified.

In the real estate business, a CRM system helps improve the relationship between the business and its customers and it helps companies know their customers' needs, interests, and buying behaviors. This enables the company to offer the right products to its customers and enables the company to provide after-sales service to its clients. For long-term customer relationships, it will increase customer loyalty and reduce customer loss. Moreover, it will reduce marketing costs and increase revenue from repeat purchases or recommend acquaintances, which means more profit for the company.

2.3 Data Mining

Data mining [7] is a process involving huge data to find patterns and relationships hidden in a dataset. Currently, data mining is used in several domains such as business, management, science, and medicine. Data mining is the evolution of data storage, and interpretation from the data stored in the database can be used to extract data to find hidden knowledge. The Cross-Industry Standard Process for Data Mining (CRISP-DM) is an open standard process model. It is the most popular analytical model that describes common approaches used by data mining experts. The CRISP-DM consists of 6 steps as shown in Figure 1.

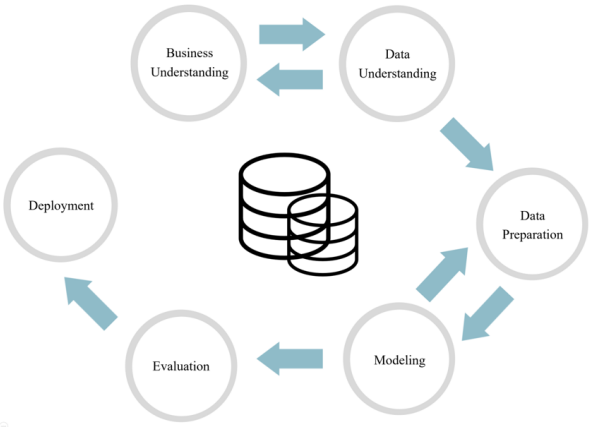


Fig.1: *The CRISP-DM process.*

1) Business Understanding

This first step focuses on determining the business objectives, assessing the business situation, determining goals, and producing a project plan.

2) Understanding the Data

This step starts with initial data collection, data description, data exploration, and also data quality verification.

3) Data Preparation

This step consists of data selection, data cleansing, data construction, data integration, and data standardization.

4) Modeling

This step is an analytical step using recommended data mining techniques, such as classification or segmentation. In this stage, several techniques can be

used to get the most accurate result. The process may need to go back to a previous step to transform data to suit each technique. Examples of techniques for analyzing data are Clustering, Association Rules, Regression, and Classification. This step consists of modeling technique selection, test design generation, model building, and model assessment.

5) Evaluation

In order to deploy the modeling results, it is necessary to measure the performance of the model to ensure it meets the objectives set in the first step and that it is reliable. This step consists of results evaluation, process review, and creating possible action lists.

6) Deployment

This final step consists of deployment planning, monitoring, maintenance planning, final report production, and project reviews.

2.4 Related Research

Theoretically, data mining has been applied to the business domain for decades. Ng and Liu [20] proposed an integrated full system that composes various data mining techniques such as inductive feature selection, deviation analysis, multiple concept-level association rules to form an intuitive understanding, and gauging customer loyalty and predicting their likelihood of defection. This paper is one of the first to apply CRM data mining in our problem domain and is used as a basis for this research henceforth.

Cheng et al. [8] have studied the Big Data Assisted Customer Analysis and Advertising architecture (BDCAA) for the real estate business. The architecture consists of three stages: 1) user 360-degree portrait and user segmentation, 2) potential customer mining, and 3) precise advertising delivery. The objective of this study is to improve the efficiency of advertisement delivery in the real estate industry. This study showed that the architecture could reach a high advertising arrival rate. It takes advantage of big data and improves real estate advertising efficiency.

Xu [9] proposed prediction of residential real estate prices based on a Back-Propagation Neural Network (BPNN). The study brought up a mixed optimizing model based on Improved Particle Swarm Optimization and BPNN (IPSO-BPNN). The study proposed a simple and feasible prediction model based on grey correlation theory and IPSO-BPNN. It detected the main factors among several inputs that influence the house price, then used the IPSO-BPNN to predict the trend. This study showed that the prediction is in accordance with the residential property market price in Changsha. Moreover, the model has advantages of good output stability, high speed of convergence, and high precision of prediction.

Yang [10] proposed research related to CRM of real estate enterprise. A model of real estate in China was constructed by studying e-commerce in order to expand management. There were several problems and risks with CRM operations, which lead to new CRM methodology. The technique used in this research is association rules mining in order to find hidden relationships in the data. It extracts the general object of the transaction which addresses the sales data. Thus, the model helps firms in marketing and advertising planning.

Ziafat and Shakerihas [11] studied the use of data mining techniques in customer segmentation. The study had initially proposed the technical integration and capabilities on business needs in order to provide useful patterns of information. The data mining technique used in this research is a clustering method used to do the segmentation. By analogy, this article addressed the blending of CRM theory and the two-step specialization technique for client segmentation.

3. RESEARCH METHODOLOGY

The overall research methodology is shown in Figure 2. Each research step follows the CRISP-DM standard, which will be described in the following subsections.

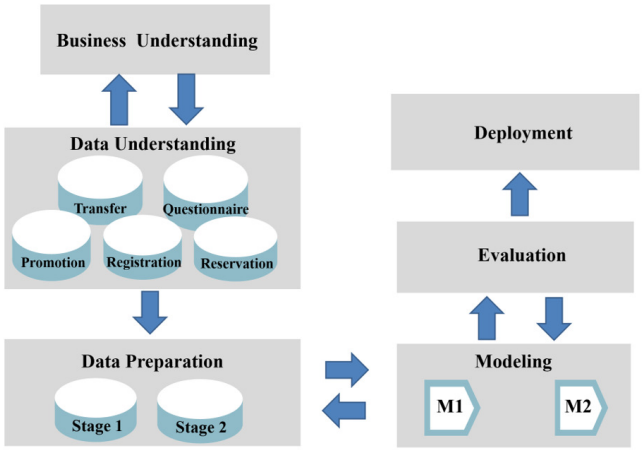


Fig.2: *Overall research methodology.*

3.1 Business Understanding

Although there are several business processes in the real estate domain, this study focuses on the selling process. For a real estate company, the products sold are homes or residences which have very high cost. These sales involve other related institutes such as banking, the Department of Land, etc. In the customer viewpoint, there are various factors which influence their decision, including lifestyle, the number of family members, location, ability to get a loan, etc. A house is a more specialized product than others because people hardly ever buy an additional house in their lifetime. Thus, if the firm cannot determine the true prospective customers who have a demand and

an ability to buy the product, they will have a lost opportunity.

The process of selling houses is divided into several parts. Each process will vary depending on the completeness of the home sold. There are three main processes, which are visitation, reservation, and transfer, as shown in Figure 3.

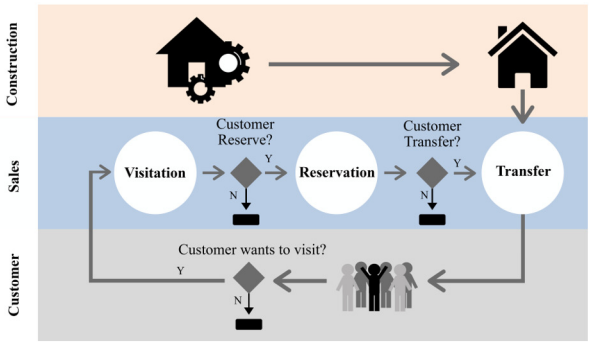


Fig.3: *Sales process of real estate business.*

Generally, the real estate firm segments its customers into several stages, which consist of prospect, potential, and purchased, as shown in Figure 4.

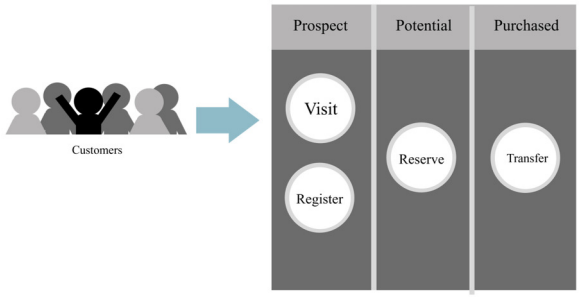


Fig.4: *Customer stages.*

3.2 Data Understanding

The data used in this work was selected from a real estate company between 2013 and 2019. There are five entities: customer questionnaires, registration data, reservation data, transfer data, and promotion data. The input data for each stage is different. The first prediction stage aims to analyze the opportunity of prospective customers to make a reservation. The required data for this stage includes customer questionnaires, reservations, and visitation data. The second stage aims to analyze customer purchasing for customers who have reserved the house and transferred ownership. The required data in this stage consists of customer questionnaires, transfer data, and promotion data. In summary, as shown in Figure 5, the stage-1 data had 40,193 examples, and 3,560 examples existed for stage-2. Note that all attributes related to customer information and contacts, which are Personally Identifiable Information (PII) [21], are encrypted with the Advanced Encryption Standard (AES) [22]. This method hides the personal information completely, but preserves the information uniqueness for table joining purposes. This procedure ensures privacy and is required by EU General Data Protection Regulation (GDPR) [23].

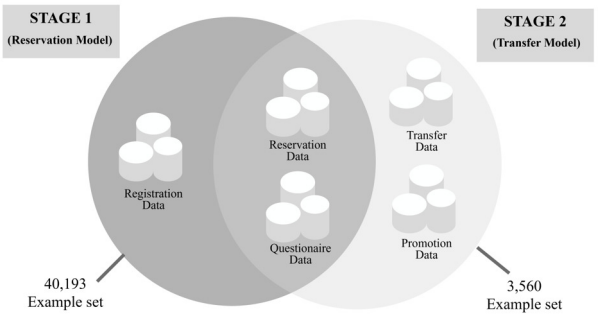


Fig.5: *Selected data sources of both stages.*

Table 1: *Data Dictionary.*

Table	Information
Registration	Project (id no.)
	Customer information (name and contact)
	Economic information (income and budget)
Reservation	Project (id no.)
	Product (id no., type, and price)
	Customer information
	(name, type, and joint customers)
	Reservation (id no. and date)
	Promotion (id no.)
Transfer	Project (id no.)
	Reservation (id no. and date)
	Product (id no., type, and base price)
	Customer information
	(name, type, and joint customers)
	Transfer (price and date)
	Promotion (id no.)
Promotion	Promotion (id no., name, price, cost, and status)
Questionnaires	Customer (id no., name, contact)
	Project (id no.)

3.3 Data Preparation

The data must be selected, identified, and prepared for modeling. It should also be cleaned and transformed to the required format for the algorithm. The steps in this process are shown in Figure 6.

1) Data mapping and integration

According to the first stage, the process starts with mapping customers' last names from registration or visitation data with customer reservation data. If their last names match, then they are related. After mapping customers by their last name, customer names and surnames are correlated with customer questionnaire data. The process flow is shown in Figure 7.

The second stage starts with the mapping of customer reservation codes with transfer data. Then it

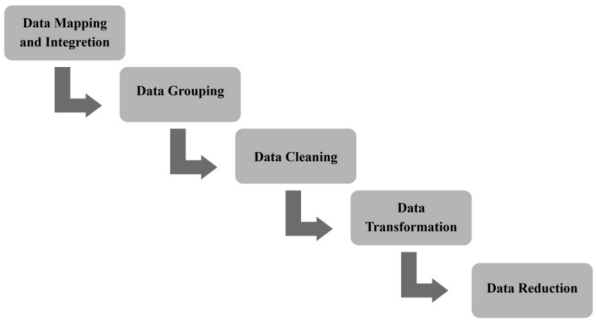


Fig.6: *Data preparation processes.*

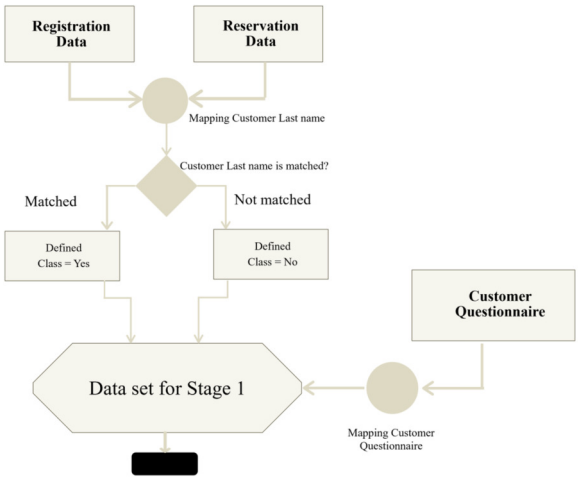


Fig.7: *Data mapping process for stage 1.*

is mapped to the customer names and surnames of data in stage 2 with customer questionnaires. The process flow is shown in Figure 8.

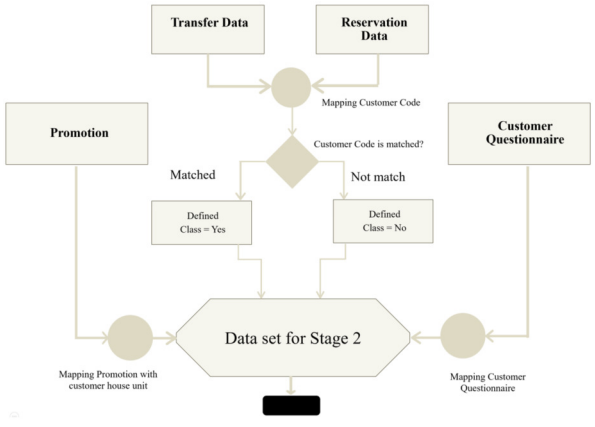


Fig.8: *Data mapping process for stage 2.*

The steps of data preparation yield the data set of 2 stages as an output, which will be transformed and cleaned in the next step.

2) Data grouping

There are several questionnaires with answer values, including interested promotions, and those are written differently but have the same meanings. Thus, the attributes that are similar should be grouped. The amount of available data after mapping of registration data, reservation data, and transfer data is shown in Figure 9.

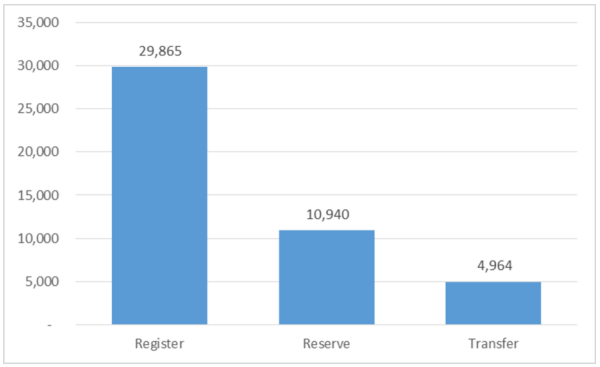


Fig.9: *Amount of examples in each stage.*

There are several questions and answers in the questionnaire data which are similar and should be grouped. The question and answer values after grouping are shown in Table 2.

Table 2: *Questionnaire Data Dictionary.*

QID	Question	Values
Q1	Age	0 - 18 / 19 - 33 / 34 - 48 / Over 48
Q2	Marital status	Single / Married
Q3	Income	50,000 - 100,000 / 100,001 - 200,000 / 200,001 – 300,000 / 300,001 - 400,000 / 400,001 - 500,000 / Over 500,000
Q4	Budget set (Million THB)	5 - 9 / 10 – 19 / 20 – 30 / Over 30
Q5	Occupation	Employee / State Enterprise / Freelance / Self-Employed / Government Agency / Housewife / Studying
Q6	Buying reason	Need a bigger house / Uncomfortable location / Separate Family / For Business / Investment / Have Children / Parking Problem / Better Environment / For descendants

3) Data cleansing

The data cleansing is done to verify and improve the data quality of the dataset since merging from several data sources with different formats often causes low-quality data issues such as inaccuracy, incompleteness, redundancy, or nonconformity. This step, theoretically, is the most time-consuming task. In this paper, the data selection process is done first. The selected fields for stage 1 are customer personal data, customer demographics, and questionnaire items. The selected fields for stage 2 are reservation code, transfer status, and customer questionnaire items. Missing values and noise are replaced. A zero replaces a missing numeric value. The ordinal attributes are replaced by the average of each

range. For example, age range is replaced by its upper bound.

4) Data transformation

In this step, the data is transformed into the proper format, which requires a custom algorithm. The attributes are transformed into one row of customer data and represented as binary. The rest of the attributes are modified as numeric values.

5) Data reduction

Data rows which lack customer information or attributes are eliminated. We also remove duplicate rows. Next, the outliers are removed after performing Exploratory Data Analysis (EDA) [12].

3.4 Modeling

After the data is prepared and ready to enter the model, the next step is to determine the best algorithm to establish the model. There are several techniques used in data mining. This study proposed two models:

Model 1: A reservation identification model for registered customers.

Model 2: A transferring identification model for reserved customers.

This research selected four predictive modeling algorithms for a comparative experiment, each of which is discussed in the following paragraphs.

1) Decision Tree (DT)

The Decision Tree algorithm [13] classifies the input data by separating each attribute in the decision node. The prediction starts from the root node by calculating the value of each attribute and then following the branches of the tree until reaching the target variable. In order to find the relationship of this attribute, the GINI Index is used. It is a measurement of the importance of each variable. The decision tree algorithm can be interpreted with minimal user intervention. It can be used for binary and multilayer classification problems since fast algorithms are known. The demonstration of the Decision Tree is shown in Figure 10.

2) Artificial Neural Network (ANN)

The key to the Artificial Neural Network [14] is attempting to imitate the human brain. The main components are input or values sent to the neuron via edges, the weight of each edge, bias, and output nodes. The Multi-layer Perceptron, as shown in Figure 11, is the most popular neural network architecture and contains intermediate hidden layers. The ANN tolerates erroneous instructions and usually achieves high prediction accuracy. It also works well with voice recognition, handwriting, or images. However, the ANN takes a long time to train a model, and the result of weighting is not interpretable.

3) Support Vector Machine (SVM)

The principle of SVM [15] is to map vector-based input into n-dimensions. For example, many results are in 2D and 3D spaces. Then a hyperplane is created to separates the input vector into different types. The SVM's dominant feature is to map the vector of input space to the feature space by the kernel. Examples of kernel functions are polynomial and radial. In the feature space, the input vector can be classified by a hyperplane. The selected kernel function of SVM can shift the data to a lower dimension to give a higher degree of linearity. Regularization and kernel choices occur only in a way so that SVM shifts the issue of parameter tuning to fit the pattern selection. The kernel model may be more sensitive than the model selection criteria.

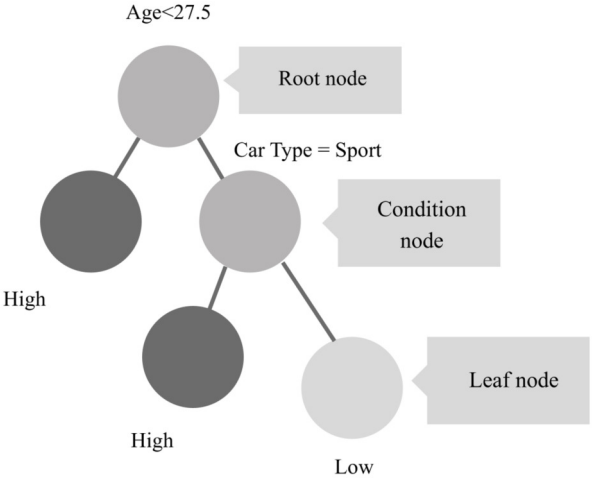


Fig.10: *Decision tree model example.*

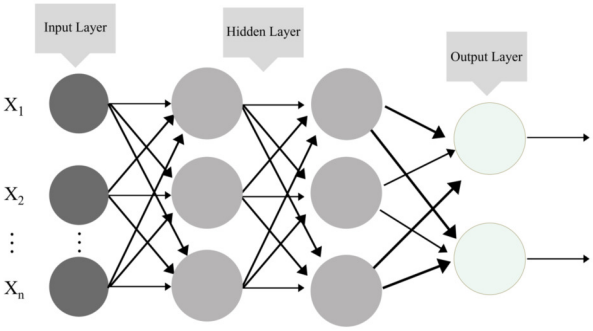


Fig.11: *ANN with multi-layer perceptron.*

4) Gradient Boosted Trees (GBT)

Gradient Boosted Trees [16] models, as illustrated in Figure 20, consist of a set of regression or classification trees. It learns advance learning sets that get predicted results through better estimation. Boosting is a flexible nonlinear regression that improves tree accuracy by dealing with weak data that is difficult to classify. Set of decision trees are made to create weak predictive sets. Although GBT gains increasing accuracy, it reduces the training speed and ability of human interpretation.

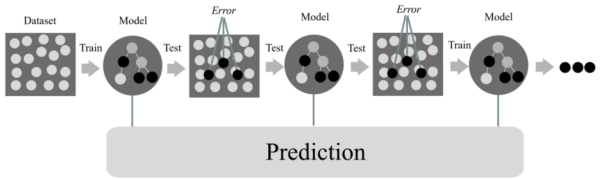


Fig.12: *Gradient Boosted Trees Example.*

3.5 Evaluation

This research employs k-fold cross-validation [17] as an evaluation method. The basis of the cross-validation method is resampling, starting with the division of data into parts, and sampling some data from that series to cross-check. Cross-accessing results are often used as an option for determining accuracy of network models, such as network architecture and classification models. This repetitive hold-out method aims to measure how much the trained model is overfitting by examining the performance. For example, the overfitting model is proven by the low resulting performance. The k value is an adjustable parameter that refers to the number of partitions (or fold). The demonstration of this method is shown in Figure 13.

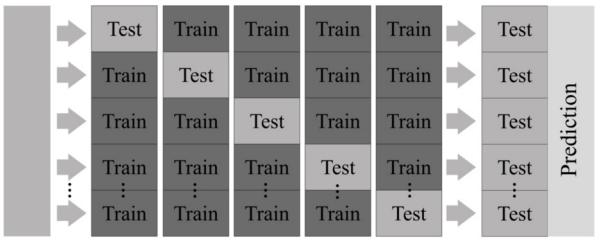


Fig.13: *K-fold cross-validation.*

3.6 Setting up of experiment

In order to gain the most accurate model, evaluated by a 10-fold cross-validation method, each algorithm optimizes its parameters by using the Evolutionary Optimization algorithm [18]. The evolutionary algorithm is a generic population-based metaheuristic optimization algorithm. It uses a process motivated by biological evolution. Candidate solutions to the optimization problem play the role of individuals in a population, and the fitness function determines the quality of the solutions.

A set of parameters for each predictive model is shown in Table 3.

4. RESULTS AND DISCUSSION

As stated in the previous section, the comparative experimental result is shown in Table 4 and illustrated in Figure 14. Each model of each stage is reported with its optimum parameter set and four

Table 3: *Parameters set for optimization.*

Algorithm	Parameter
Decision Tree (DT)	Maximal depth (*MaxDepth*) Minimum confidence (*MinConf*)
Artificial Neural Network (ANN)	Number of hidden layers (*HidLayer*) Number of hidden nodes (*HidNode*) Learning rate (*LR*) Momentum (*MT*)
Support Vector Machine (SVM)	*Gamma* *nu*
Gradient Boosted Tree (GBT)	Number of trees (*NumTree*) Maximal depth (*MaxDepth*) Minimum confidence(*MinConf*)

measurements: accuracy (Acc), precision (Pre), recall (Rec), and F-measure (F). The accuracy is calculated by the ratio between correctly classified samples over all of the training set. It implies the number of correctly classified instances over all training data. The precision refers to the ratio of number of true positive instances by number of predicted positive instances. The recall, also known as sensitivity, is the ratio of number of true positives over actual positive ones. The F-measure is the harmonic mean of precision and recall. In brief, the accuracy focuses on overall data including positive and negative values, while the F-measure does not take the negative values into account. It focuses on positive values (confirmation to reserve or confirmation to transfer). Therefore, the results reported in Figure 14 are compared the accuracy and F-measure because the precision and recall are reflected with the F-measure.

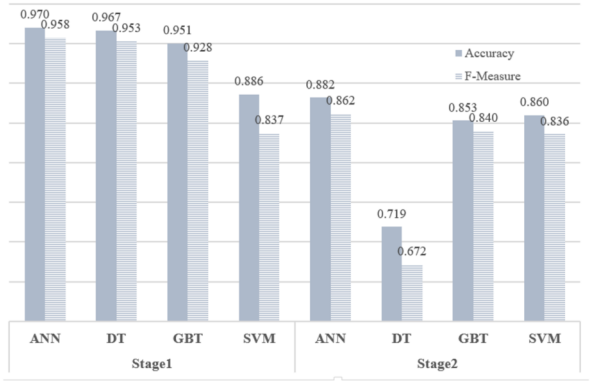


Fig.14: *Comparative experimental results.*

As shown in Table 4, the most accurate model for both stages is the ANN, with 0.970 and 0.882 accuracy respectively. In the same fashion, the F-measure for both stages of the ANN are also superior to others, with values of 0.958 and 0.862. This means the ANN is suitable for both the customer identification and loyalty discrimination tasks. As stated in the previous section, the reported parameter sets are optimized to obtain the best model. The optimized number of hidden layers of both models is one layer.

This result reflects the theoretical assumption about the ability of ANN to handle complex data and problems. The single hidden-layer architecture of ANN is enough to handle non-high dimensionality problems. When comparing with a baseline method like the Decision Tree (DT), the accuracy is not significantly different in stage 1 (0.970 and 0.967), but quite obviously is significant in stage 2 (0.882 and 0.719). The F-measure of those methods have the same trend.

The parameter set of each model is quite interesting. When comparing DT and GBT, the optimum maximal depth of DT is higher than GBT since the GBT provides multiple regression trees based on Gradient Descent algorithm. Thus, the complexity is reflected by the number of trees instead. On the other hand, when comparing ANN with SVM, the parametric method like ANN is more suitable than the non-parametric method like SVM. The reported performance was measured by 10-fold cross validation which aims to prevent the overfitting problem. However, additional dataset should be used to test these models in real production environment.

Table 4: *Experimental results.*

Stage	Model	Acc	Pre	Rec	F
1	DT *MaxDepth:* 20 *MinConf:* 0.01	0.9667	0.9931	0.9157	0.9528
	ANN *HidLayer:* 1 *HidNodes:* 40 *LR:* 0.20 *MT:* 0.20	0.9700	0.9925	0.9254	0.9578
	GBT *NumTree:* 20 *MaxDepth:* 5 *MinConf:* 0.24	0.9506	0.9903	0.8739	0.9284
	SVM *Gamma:* 100 *nu:* 0.20	0.8860	0.8792	0.7986	0.8370
2	DT *MaxDepth:* 13 *MinConf:* 0.25	0.7191	0.7152	0.6330	0.6716
	ANN *HidLayer:* 1 *HidNodes:* 30 *LR:* 0.20 *MT:* 0.30	0.8821	0.9227	0.8080	0.8616
	GBT *NumTree:* 32 *MaxDepth:* 7	0.8530	0.8314	0.8481	0.8396
	SVM *Gamma:* 97 *nu:* 0.23	0.8600	0.8931	0.7857	0.8359

Although their uninterpretable results cause the black-box phenomena in model usage when compared with tree-based or rule-based methods like Decision Tree or Gradient Boosted Tree, the highest predictive performance of ANN leads to the model deployment embedded with a web application to predict the chance of reservation in stage 1, and transferring in stage 2.

The proposed method is more novel in the perspective of CRM than the existing works [8] [9] [10] [11] [20] since it captures the customer status along their journey instead of predicting the single-stage indicator or status. Moreover, in terms of prediction, the results for accuracy and F-measure are proper and better indicators for the model performance, when compared with.

Moreover, the authors present the statistical analysis results for the stochastic algorithm. The kind of test selected is a two-tailed t-test for accuracy [19] as shown in Table 5. The statistical test proved that the ANN is the most accurate model for both stages. However, other models should be subjected to further in-depth analysis to conclude their ranks.

Table 5: *Statistical analysis with two-tailed t-test.*

Stage	Algorithm		p-value	Conclusion
	First	Second		
1	ANN	GBT	0.0099*	**ANN > GBT**
	ANN	DT	0.0093*	**ANN > DT**
	ANN	SVM	0.0023*	**ANN > SVM**
	GBT	DT	0.2780	DT > GBT
	GBT	SVM	0.2121	GBT > SVM
	DT	SVM	0.7945	DT > SVM
2	ANN	GBT	0.0098*	**ANN > GBT**
	ANN	DT	0.0090*	**ANN > DT**
	ANN	SVM	0.0027*	**ANN > SVM**
	GBT	DT	0.0015*	**GBT > DT**
	GBT	SVM	0.3027	GBT > SVM
	DT	SVM	0.0013*	**SVM > DT**
	ANN	GBT	0.0098*	**ANN > GBT**

Overall accuracy ranking:
$\mu_{\text{Stage1(ANN)}} > \mu_{\text{Stage1(DT)}} > \mu_{\text{Stage1(GBT)}} > \mu_{\text{Stage1(SVM)}}$ and
$\mu_{\text{Stage2(ANN)}} > \mu_{\text{Stage2(GBT)}} > \mu_{\text{Stage1(SVM)}} > \mu_{\text{Stage1(DT)}}$
* Significant different at $\alpha = 0.05$

The application is embedded in the current business process. It assists the salespersons and other stakeholders in determining which potential customer will reserve a single house product and the probability of achieving the transfer stage. To demonstrate the deployed application, Figures 15 to 19 show the information flow of the application and its example results in four different states as summarized in Table 6.

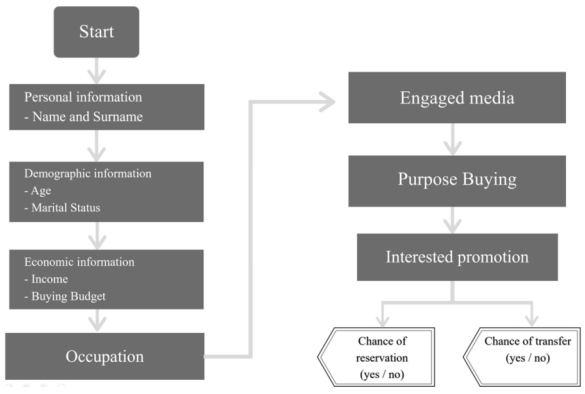


Fig.15: *Information flow of the developed application.*

Table 6: *Output scenario of the developed applica-tion.*

Case no.	State
1	Not reserve / Not transfer
2	Reserve / Transfer
3	Reserve / Not transfer
4	Not reserve / Transfer

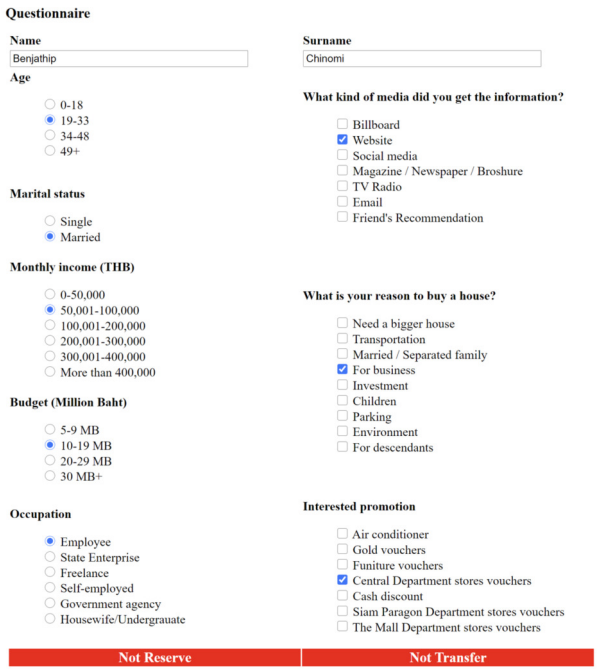


Fig.16: *The customer journey analytical application: case 1.*

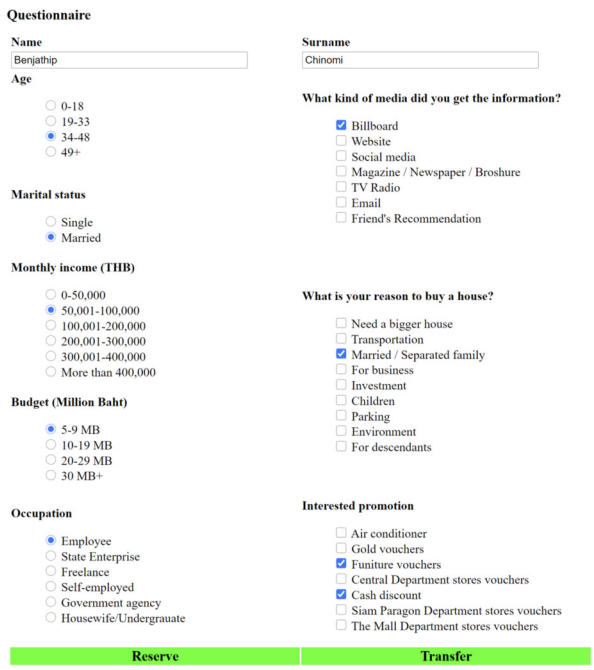


Fig.17: *The customer journey analytical application: case 2.*

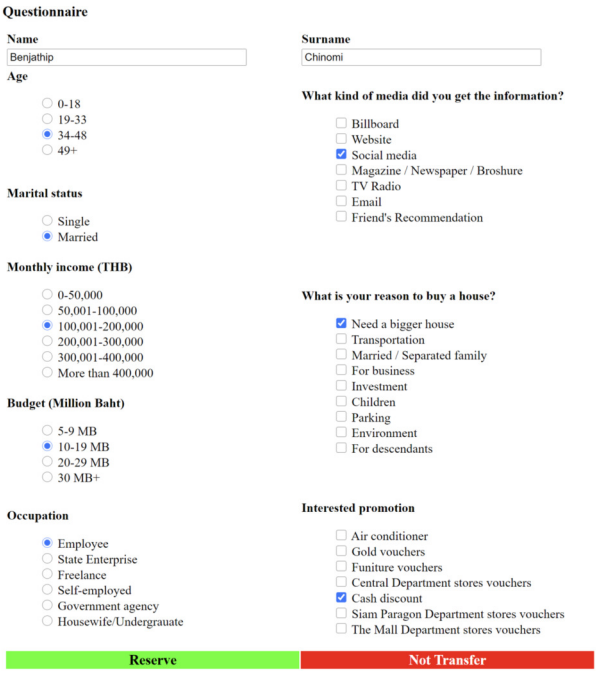


Fig.18: *The customer journey analytical application: case 3.*

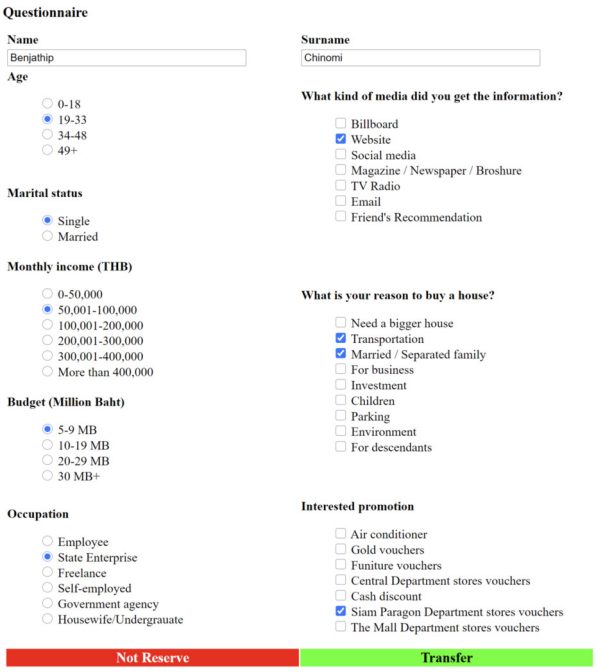


Fig.19: *The customer journey analytical application: case 4.*

5. CONCLUSIONS

Due to rapidly growing competition in the real estate domain, sellers need data analytic tools to identify potential customers along their journey. This research studies the characteristics and business model of real estate with a preliminary study of the problems and background of real estate companies and trends. The concept of Customer Relationship Management (CRM) is described to show the impacts of customer relationships. This research proposed a two-stage customer retention model for real estate firms, specifically for single house products. The research methodology is based on the CRISP-DM scheme. It started with understanding the problem and transforming it into a data analysis solution. Then the researchers collected the necessary data from different sources. The first stage combined the registration data, reservation data, and customer questionnaire answer. The second stage conformed to stage 1 and also included promotion data and excluded registration data. Then, the researchers performed conversion and integration of the raw data into an analysable format. There are five steps in order to prepare data before modeling: data mapping and integration, data grouping, data cleansing, data transformation, and data reduction. Subsequently, the data was prepared; there are two models which are for single houses in stage 1 and 2. Four classification techniques were used to get the best model. The ANN with 1 hidden layer presented the highest accuracy and F-measure of 0.970 and 0.958 in stage 1 and 0.882 and 0.862 in stage 2, compared to other algorithms. In deployment, the best ANN model was selected to establish the retention prototype which can determine customer class in two stages. This research highlights how data mining techniques can help the sellers to seek potential customers and improve the CRM objectives.

One limitation of this work is that the training data set was not large. That is caused from the lack of a quality data-entry process and can lead to overfitting problems. Some other algorithm may be further employed to empower the prediction system. Moreover, the data used in this research is based on an existing relational database but should be expanded to include more data such as social media contents and geographical information.

ACKNOWLEDGEMENTS

We extend our sincere thanks to all individuals and firms who contributed data and motivated us while we did this research work.

References

[1] CBRE (Thailand) Co., Ltd., "Thailand Real Estate Market Outlook 2019," CBRE. [Online]. Available: `https://www.cbre.com/research-and-reports/Thailand-Real-Estate-Market-Outlook-2019`. [Accessed: 30-Jun-2019].

[2] A. B. Galvao and K. Sato, "Human-Centered System Architecture: A Framework for Interpreting and Applying User Needs," *Volume 3a: 16th International Conference on Design Theory and Methodology*, 2004.

[3] Kreyon, "10 Ways Business Process Automation is changing Real Estate," *kreyon systems | Blog | Software Company | Software Development | Software Design*, 16-Sep-2019. [Online]. Available: `https://www.kreyonsystems.com/Blog/10-ways-business-process-automation-is-changing-real-estate/`. [Accessed: 30-Apr-2019].

[4] D. E. Holmes, Data mining. Berlin: Springer, 2012.

[5] Colliers International, "Thailand Property Research Reports Q3 2019," Thailand Property Research Reports Q1 2019 | Thailand | Colliers International. [Online]. Available: `https://www.colliers.com/en-th/thailand/insights/quarterly-reports`. [Accessed: 01-Oct-2019].

[6] N. Bhatnagar, "Customer Relationship Marketing: Customer-Centric Processes for Engendering Customer- Firm Bonds and Optimizing Long-Term Customer Value," *Advances in Customer Relationship Management*, Nov. 2012.

[7] K. J. Cios, *Data mining: a knowledge discovery approach*. New York, NY: Springer, 2010.

[8] X. Cheng, M. Yuan, L. Xu, T. Zhang, Y. Jia, C. Cheng, and W. Chen, "Big data assisted customer analysis and advertising architecture for real estate," *2016 16th International Symposium on Communications and Information Technologies (ISCIT)*, 2016.

[9] H. Xue, "The Prediction on Residential Real Estate Price Based on BPNN," *2015 8th International Conference on Intelligent Computation Technology and Automation (ICICTA)*, 2015.

[10] N. Yang, "Research on the customer relationship management of real estate enterprise," *2010 IEEE 2nd Symposium on Web Society*, 2010.

[11] H. Ziafat and M. Shakeri, "Using data mining techniques in customer segmentation," *Int. Journal of Engineering Research and Applications*, vol. 4, no. 9, pp. 70–79, 2014.

[12] F. Hartwig and B. E. Dearing, *Exploratory data analysis*, Newbury Park: Sage Publ., 1994.

[13] K. Singh, "The Comparison of Various Decision Tree Algorithms for Data Analysis," *International Journal of Engineering and Computer Science*, pp. 21557–21562, Jan. 2017.

[14] B. Liu, Y. Wei, Y. Zhang, and Q. Yang, "Deep Neural Networks for High Dimension, Low Sample Size Data," *Proceedings of the Twenty-Sixth*

International Joint Conference on Artificial Intelligence, 2017.

[15] L. Hamel, Knowledge discovery with support vector machines. Hoboken: John Wiley & Sons, 2009.

[16] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," *Frontiers in Neurorobotics*, vol. 7, 2013.

[17] K. Ramasubramanian and A. Singh, "Machine Learning Model Evaluation," *Machine Learning Using R*, pp. 425–464, 2016.

[18] D. Simon, *Evolutionary optimization algorithms*, Chichester: Wiley-Blackwell, 2013.

[19] F. M. Dekking, A modern introduction to probability and statistics: understanding why and how. London: Springer, 2010.

[20] K. Ng and H. Liu, "Customer Retention via Data Mining," *Artificial Intelligence Review*, vol. 14, pp. 569-590, 2000.

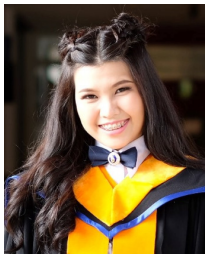
[21] P. M. Schwartz and D. J. Solove, "Reconciling Personal Information in the United States and European Union," *SSRN Electronic Journal*, pp. 877–916, 2013.

[22] A. Bogdanov, D. Khovratovich, and C. Rechberger, "Biclique Cryptanalysis of the Full AES," *Cryptology ePrint*, 2011.

[23] P. Voigt and A. von dem Bussche, *The EU General Data Protection Regulation (GDPR): a practical guide*, Cham, Switzerland: Springer, 2017.

[24] Thai government, "Personal Data Protection Act", 2019.

Sotarat Thammaboosadee was born in 1982 and received B.Eng. degree in Computer Engineering in 2003, and M.Sc. degree in Technology of Information System Management in 2005 from Mahidol University, Thailand. In 2013, he received Ph.D. degree in Information Technology from King Mongkut's University of Technology Thonburi, Thailand. He is now an assistant professor at IT Management Division, Faculty of Engineering, Mahidol University, Thailand. He is also the director of Datalent Team, Data Talent Development Research Group. His research interests include Data Science, Data Privacy, Data Governance, Data Stewardship, and Applications of Data Mining on Several Domains, such as legal, economic, and healthcare domains.



Benjathip Chinomi received B.Sc. degree in Information and Communication Technology from Mahidol University, Thailand in 2015, and M.Sc. degree in Information Technology Management from Mahidol University, Thailand in 2017. Her research interests include Data Mining, and Customer Analytics.

Ehab K. A. Mohamed is the Dean and Professor of Accounting at the Faculty of Management Technology, German University in Cairo. Prior to joining GUC he worked for 10 years at Sultan Qaboos University, Oman. He graduated from Cairo University and received his M.Sc. & Ph.D. from Cass Business School, London. He is a Fellow of the Chartered Institute of Internal Auditors, UK. His areas of research are in auditing, fraud, performance measurement, business education, financial reporting, and corporate governance. He has published a number of papers in international refereed journals and presented papers at numerous international conferences. He published two books on financial accounting and auditing.