Sites: https://www.jurnal-umbuton.ac.id/index.php/Pencerah
DOI: https://doi.org/10.35326/pencerah.v7i2.1126

Open Access

SANG PENCERAH
Jurnal Ilmiah Universitas Muhammadiyah Buton

E-ISSN: 2655-2906, P-ISSN: 2460-5697 Volume 7, No 2, Tahun 2021

Real Estate Segmentation: A Model of Real Estate Decision Support System

Genesis Sembiring Depari

Universitas Pelita Harapan, Medan Campus, Indonesia
Korespondensi: genesis.sembiring@uph.edu

Info Artikel

Diterima *1 Mei 2021*

Disetujui *19 Mei 2021*

Dipublikasikan *21 Mei 2021*

Keywords:
Real Estate Segmentation, K-Means Algorithm, X-means Algorithm, K-medoid Algorithm, Random Forest

Abstract

Due to human limitations of computational thinking, the quality of rational decision-making is constrained, and as a result, people encounter bounded rationality. A decision support system is widely demanding in tackling this problem, especially in real estate management. This study focuses on 3 main purposes. Firstly, comparing K-means, X-means and K-medoid algorithm performance in clustering sold house characteristics to be further used for pricing houses. Second, characterizing each cluster for developing a suitable marketing strategy by utilizing machine learning technology. Lastly, providing a managerial implication as a decision support system for assisting stakeholders in making a decision. Eventually, K-means and X-means algorithm show very similar performance. X-means can automatically determine the number of clusters while k-means utilize the elbow method to find the optimum number of clusters. Three clusters were identified as cluster 0, cluster 1, and cluster 2. Cluster 0 was occupied by 85.77% of low house prices. There are two practical implications of this study. Firstly, the results of clustering analysis which reflected in a model of decision support system. Second, an intuitive and comprehensive methodological framework is presented for helping stakeholders designing a decision support system.

1. Background

Simon (1957) contends that the human limits in computational thinking constrain the quality of rational decisions; therefore, people experience bounded rationality. Besides, recent researches show that heuristics decision-making has led to a judgmental error in valuation prediction (Tidwell, 2011). Moreover, Geltner (Geltner, 1989) stated that real estate appraisers often only rely on previous value judgment when dealing with uncertainty. This also often leads people to perform a problem of heuristics in making decisions. Therefore, a reliable decision support system is crucially notable for decision-makers.

For a real estate appraiser, it is essential to have an informative system to identify the similar characteristics of houses before estimating their value. Besides, a real estate agent or broker also needs this information for designing the next possible marketing strategy and product development. Moreover, Wu and Sharma (C. Wu & Sharma, 2012) contend submarket segmentations of houses are also crucial for mortgage leaders, house developers, government, non-profit organizations, individual homeowners for issuing a wise and appropriate decision into the market. This phenomenon explains the importance of a decision support system for avoiding a subjective decision.

In utilizing real estate data for appraisal purposes, many approaches have been built for many years such as neural networks (Peterson & Flanagan, 2009), random forest (Antipov & Pokryshevskaya, 2012), fuzzy system (Kuşan et al., 2010), etc. Moreover, data-driven methods are more intuitive for dynamic and flexible house prices (C. Wu & Sharma, 2012). Therefore, data-driven methods along with computational power are promising techniques to provide a decision support system. Besides, in predicting house prices, there have been also many studies such as using deep learning and random forest to predict house prices (de Aquino Afonso et al., 2020), using linear regression to predict the housing price (L. Wu & Brynjolfsson, 2015) performing artificial neural networks to predict the housing price (Selim, 2009), etc. Furthermore, for clustering houses, many studies also have been conducted such as segmenting the housing market in Seoul (B. T. Kim, 2000), analyzing the house price determinants through a cluster analysis (Kang, 1995), etc.

This study was intended to fulfill the gap of housing segmentation literature on providing a decision support system of analyzing the characteristics of clusters by leveraging 21,613 sold houses data. K-means, X-mean, and K-medoid algorithms are compared for grouping the similarity of sold houses based on previous transaction data into several groups. To understand the characteristics of each cluster, a classification task and variable weighting were performed by utilizing random forest, grid search algorithm, and 10-fold cross-validation technique. Eventually, a decision support system is proposed to help decision-makers design suitable strategies for the market.

2. Literature Review

Clustering algorithm has been widely used in many business applications such as clustering employee profile for providing a suitable training program for the employee (Esmaeilzadeh et al., 2016), clustering operational managerial research (Brusco et al., 2017), clustering online learning resources (Q. Wu et al., 2016), clustering customer behavior of bank's customers (Abbasimehr & Shabani, 2019), investigating user's search experience and satisfaction (Burt & Liew, 2012), clustering railway driving mission (Yatchev et al., 2012), evaluating the use of intellectual intelligence tools (Fourati-Jamoussi et al., 2018), etc.

Submarket houses was delianated using a modified data-driven framework considering spatial heterogeneity (C. Wu et al., 2018). The result shows that in delineating the submarket of houses the model works better than a traditional framework. Moreover, Chhetri et al (Chhetri et al., 2009) performed a spatial autocorrelation combined with multivariate analysis to analyze the spatial pattern of house prices in the Brisbane metropolitan area. However, the previous research

results provide evidence that the submarket of houses needs to be considered for having a better prediction accuracy.

Kim (B. T. Kim, 2000) segments the housing market in Seoul using the size of condominium, rent, and price as input attributes. Still in Seoul, Kang (Kang, 1995) analyze the house price determinants through cluster analysis. Moreover, Kim and Park (G. S. Kim & Park, 2003) leverage the hedonic pricing model to analyze the price differences among several districts in South Korea. From several studies about house segmentations in South Korean, there is no research focuses on analyzing the characteristic of each cluster so that can be used as a decision support system.

In the United Kingdom, Hoesli, Lizieri, and MacGregor (Hoesli et al., 1997) studied the United Kingdom Commercial Property market segmentation. The results show that property type was found as an important attribute for distinguishing property market behavior. Discriminant analysis and cluster composition stability testing were also used to establish the significance of the findings. Eventually, it was determined that the research conclusions were consistent with the results. However, as previously reported, this study still lacks a straightforward explanation of the characteristics of each generated cluster, making it impossible to use as a decision support method.

3. Data and Methodology

Table 1. Data Description

Variables	Description
price	house prices
bedrooms	number of bedrooms per house
bathrooms	number of bathrooms per house
sqft_living	square footage of the house
sqft_lot	square footage of the lot
floors	total floors in a house
waterfront	House which has a view to the waterfront
view	Has been viewed
condition	The condition of the houses
grade	the grade was given by king county grading system
sqft_above	square footage of house apart from the basement
sqft_basement	square footage of house from basement
yr_built	built year
yr_renovated	renovated year
zipcode	zipcode
lat	latitude coordinate
long	longitude coordinate
sqft_living15	living room area in 2015
sqft_lot15	lot size area in 2015

The dataset was retrieved from Kaggle open-source datasets collected from housing sales in King County, USA. The dataset consists of 19 selected attributes with 21,613 sold house observations from May 2014 to May 2015. The dataset is completely described in table 1. King County is located in Washington City, the

largest county under Washington City with more than 2 Million population and approximately 893,157-unit houses based on Census in 2016.

Based on the data described in table 1, 19 sold house attributes are available, which are a potential insight to be further analyzed by real estate entrepreneurs for grouping houses based on their similarity. To have this information, K-means, X-mean, and K-medoid algorithms are compared for classifying the similarity of real estate based on previous sold houses data into several groups.

The idea of K-means was first introduced by MacQueen (MacQueen, 1967) and further developed again as a standard algorithm by Lloyd (Lloyd, 1982). K-means clustering algorithm aims to group data that are similar to each other and is often called unsupervised machine learning since it does not need a data label. K-means algorithm regulates a series of k clusters and assigns each of the data to be in one particular cluster. The clusters contain similar data which is determined by the distance calculation between them. K-means algorithm works through several steps as follows. Firstly, input the random number of clusters. This can be started from 2-9 clusters. The process was iterated several times to have average within-cluster distance which was further used to select the best number of clusters. Secondly, a centroid is calculated for each cluster. Third, the distance between centroid is calculated, and creating a group based on minimum distance. For better understanding, K-means clustering can be formulated as follow. Let $Z=\{z_i\}$, i = 1….., n represent the n-dimensional data to be segmented into a set of M clusters, $C = \{c_m , m = 1,….., M\}$. As mentioned earlier, the K-means algorithm seeks to find the minimized squared error between average points of clusters. If μ_m is the averages point of cluster c_m. Therefore, a squared error between average points in cluster c_m can be formulated as follow.

$$J(c_m) = \sum_{z_i \in c_m} ||z_i - \mu_m||^2 \dots\dots\dots\dots\dots\dots\dots\dots (1)$$

Therefore, the sum of squared error between average points of all clusters is minimized which formulated as follows.

$$J(C) = \sum_{m=1}^{M} \sum_{z_i \in c_m} ||z_i - \mu_m||^2 \dots\dots\dots\dots\dots\dots\dots (2)$$

Because the K-means algorithm needs to deal with the number of clusters, Pelleg and Moore (Pelleg & Moore, 2000) proposed the X-means algorithm which automatically determines the number of clusters by optimizing a criterion such as Bayesian Information Criterion or Akaike Information Criterion. Moreover, Meen et al (Meen et al., 2014) reduce the local optima of the traditional K-means algorithm by dynamically adjust the initial cluster center by selecting the points randomly. Besides, Kaufman and Rousseeuw (Kaufman & Rousseeuw, 2009) proposed to use the median of the data to represent a cluster instead of using the mean of the data which prior proposed by MacQueen (MacQueen, 1967). Therefore, in this study, we compared these 3 clustering algorithms to find the most suitable algorithm in dealing with house datasets.

In order to reduce the dimensionality, a PCA algorithm also was utilized. PCA is a well-known method used for reducing model dimensionality by extracting new attributes based on original attributes characteristics such as variation of the data. PCA was first introduced by Pearson (Pearson, 1901) and then built again by Hotelling in 1933. PCA has been using in many applications such as dimensionality reduction for network intrusion detection (Vasan & Surendiran, 2016), forecasting daily stock return (Zhong & Enke, 2017), mining human activity (el Moudden et al., 2016), etc.

To see the characteristics of each cluster, a classification task was performed using the random forest algorithm. A random forest algorithm was employed to predict the house prices and eventually weigh the importance of attributes for characterized each cluster which then considered clusters characteristics. The analysis process was conducted by using *Rapidminer* analytic tool. The process is shown in figure 1. Firstly, the dataset was feed into the system, then data preparation was conducted in this stage. As date and house ID are considered less related to our segmentation model, these two attributes were excluded by performing select attributes operator (figure 1). In order to reduce the computational cost, a stratified random sampling technique was utilized. The stratified random sampling technique divides the datasets into several groups that have similar characteristics. Furthermore, the representation of each group is taken by using a probability sampling technique. Besides, to have a comparable value, the attributes were normalized by using the z-transformation technique. Z-transformation technique was first introduced by Goldin and Kanellakis in 1995 and intended to transform the input vectors to output vectors in which the averages are approximately 0 and the standard deviation is close to 1. The Z-transformation technique is defined as follows.

$$Z_m = \frac{X_m - \bar{X}}{S} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (3)$$

Where, Z_m, X_m, $\bar{X}$, and S are Z-transformation, original data of sample, the sample mean and standard deviation, respectively. Afterward, to have a better explanation, the house prices are discretized into 3 groups. Discretizing prices are described in table 1. This discretizing aims to classify the house prices then easier to interpret the result of real estate segmentation. Moreover, to have this discretizing valid, a real estate expert's opinion is followed and implemented in determining the size of the range of each price discrete.

Table 2. Price Classifications

Price	Range Nominal	Absolute Count
Medium	400.001-835.000	939
Low	75.000-400.000	766
High	835.001-7.700.000	240

Furthermore, a PCA analysis was leveraged to reduce the model dimensionality. Variance with a 95% threshold was kept. Therefore, all the datasets with a cumulative variance greater than the variance threshold are removed from the ExampleSet. This parameter can be adjusted through the PCA operator which is available in *rapidminer* software. The results of PCA are three new attributes that are then used for clustering purposes. K-means, X-means, and K-medoid algorithms were performed and compared in segmenting the house sold

characteristics. Processing time and average within-cluster distance are used to compare the performance of these 3 algorithms.

To have better explanation ability, the aggregate operators in *rapidminer* are used by labeling the price of houses. This allows the algorithm to count the number of examples for each combination of cluster and label attribute (prices discretizing). Reorder attributes, sort, and rename operators were also utilized for better results visualization. The complete analysis process in *rapidminer* software is drawn in figure 1.

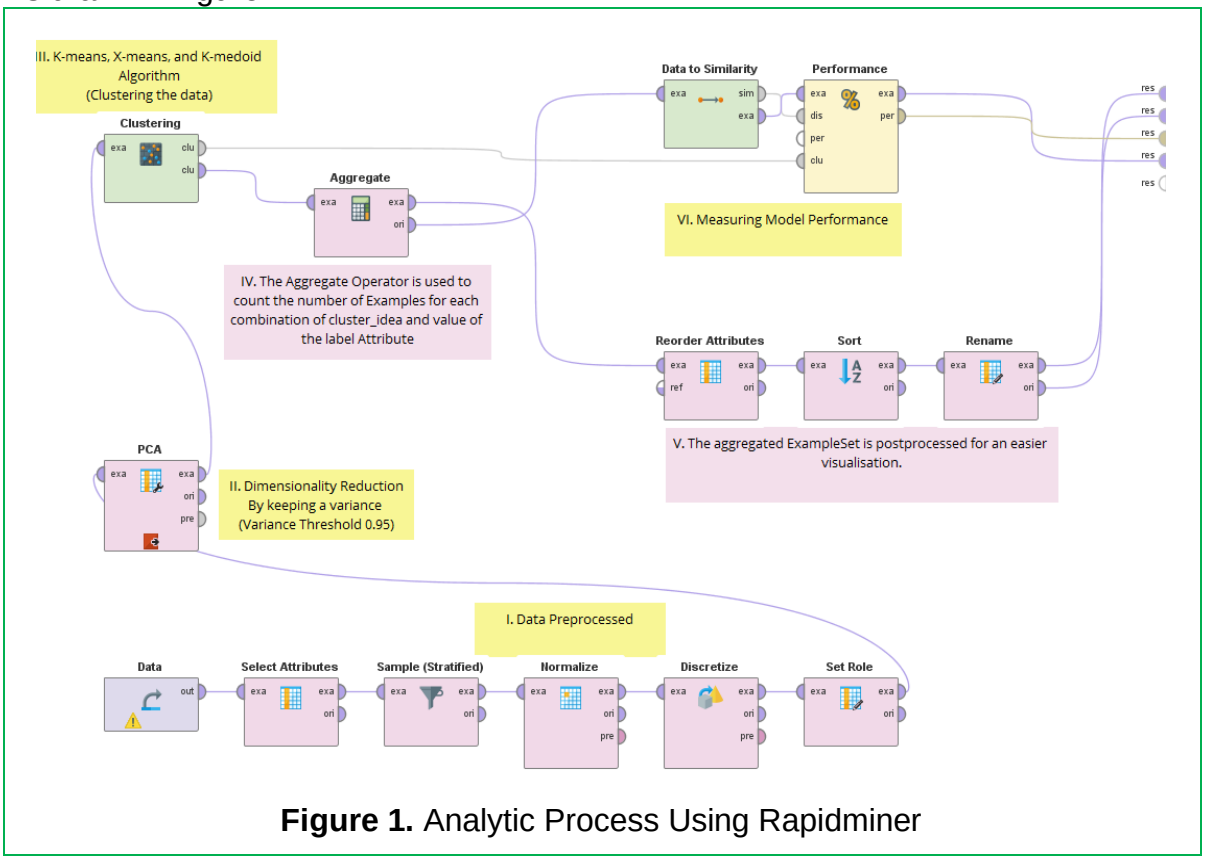

Figure 1. Analytic Process Using Rapidminer

To understand the characteristics of each cluster, a classification task using the random forest algorithm was conducted. Random forest first introduced by Ho in 1995 and then developed again by Breiman (Breiman, 2001). A random forest algorithm is used to correct overfitting in the decision tree (Hastie et al., 2009). To build a random forest classifier, first develop some decision trees (bootstrap samples) from original data, second estimating newly inserted data by averaging the estimations of decision trees previously built (majority votes for classification, averaging for regression). In order to measure the performance of the model, out-of-bag (OOB) samples are feeding into the bootstrap tree and aggregate the estimations.

To avoid the overfitting problem, a 10-fold cross-validation technique was performed. Besides, tuning the hyperparameters, a grid search technique was utilized. The number of trees which range from 11, 30, 50, 70, and 100 were examined along with several criterions such as information gain, information gain ratio, and Gini index. These parameters were randomly combined and processed. Combination with the highest accuracy is then considered the most optimum

parameter used in the prediction model. The complete classification task in *rapidminer* was drawn in figure 2.

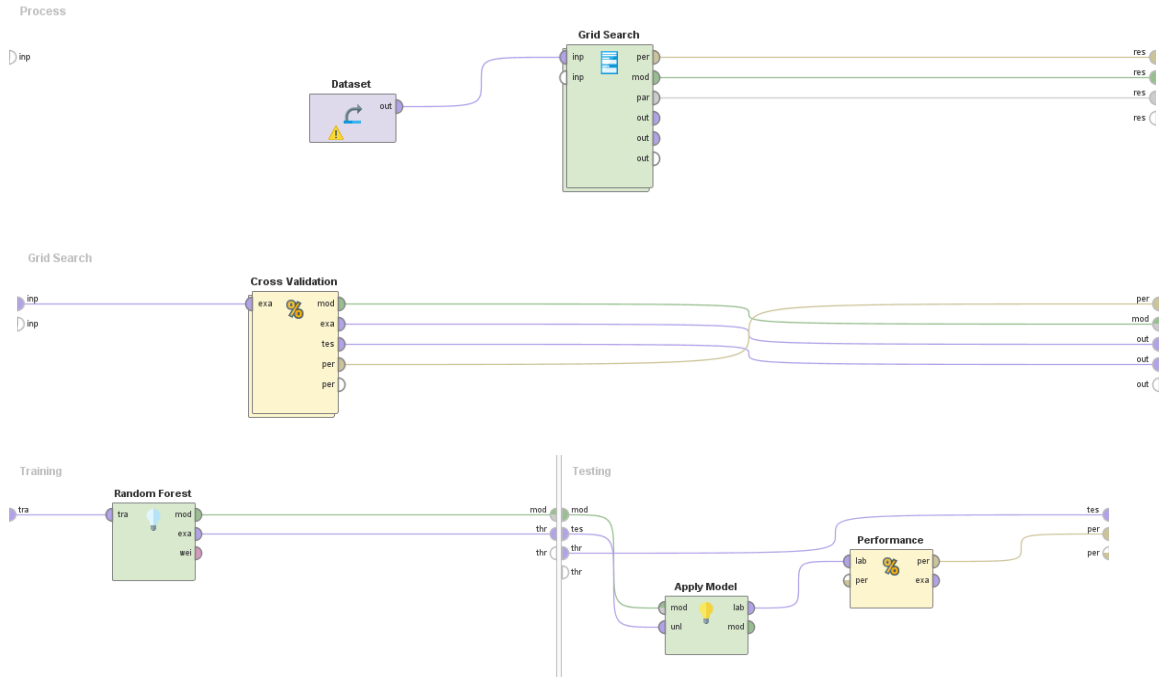

Figure 2. Classification Task Using Random Forest Algorithm

4. Results and Discussion

In the first stage, we run a data preprocessing stage. This stage encompasses 3 major analyses: data sampling using stratified random sampling, data normalization using the z-transformation technique, and price discretizing. Since we have 21,613 data observations, stratified random sampling was used to select 10% of the population. Afterward, the Z-transformation technique was performed to normalize input variables in which the averages are approximately 0 and the standard deviation is close to 1.

Normalized attributes are then used as an input for dimensionality reduction processed by Principle Component Analysis Operator. From 18 selected attributes, the dimension was reduced until only 3 principal components. The results are visualized in figure 3. To have a better visualization, the results of PCA were labeled into low, medium, and high prices. Low, medium, and high prices are green, blue, and orange colors respectively. From figure 3, it can be seen that the datasets are overlapped which means cannot be clustered by only using the sold price of houses. Therefore, a robust clustering algorithm is demanded including in determining the sufficient number of clusters. To have a broader picture regarding the result of PCA, 3 principle components are described in table 4.

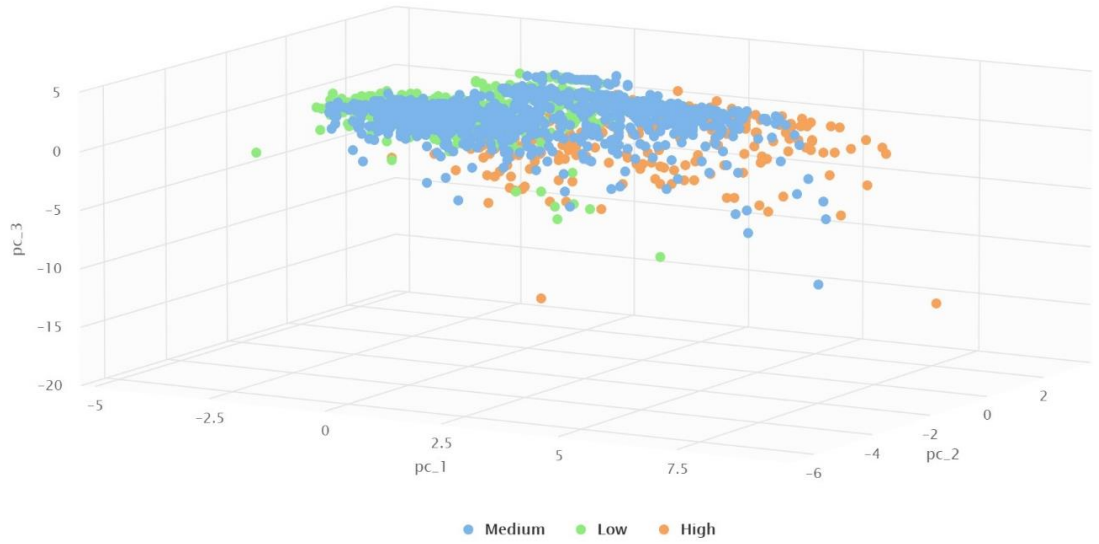

Figure 3. Principle Components Respect to House Prices

Table 4. Statistic of Principle Components

Principle Component	Min	Max	Average	Deviation
PC_1	-5.164	9.729	0	2.232
PC_2	-6.662	3.557	0	1.436
PC_3	-15.796	2.147	0	1.364

In order to cluster the similarity of attributes, three candidate clustering algorithms are examined and compared. Those algorithms are K-means, X-means, and K-medoid algorithms. The performances of these three algorithms were measured and compared in terms of processing time and average within-cluster distance. The results of the algorithm performance comparison are described in table 3. The results show that K-means and X-means reveal a very close performance on processing time and Average within-cluster distance with 14s processing time, -2494.657 and -2499.439 Average within-cluster distance respectively. However, K-medoid requires more processing time (73s) and a larger Average within-cluster distance (-6341.951). Besides, the X-means algorithm automatically determines the number of clusters and in this case, 3 clusters are developed.

Table 3. Results of algorithm comparison

Clustering Algorithm	Processing Time	Average within-cluster distance
K-means	14	2494.657
X-means	14	2499.439
K-medoid	73	6341.951

Based on the results shown in table 3, the K-means algorithm was then selected as our clustering algorithm. Furthermore, three principal components are used as input variables which are then clustered by the k-means algorithm. Before processing the three principle components for clustering purposes, the number of clusters needs to be determined. In order to determine the number of clusters, we observed the average within-cluster distance of the attributes. The average within-

cluster distance was described in table 5. From the table, we can see that the high value dropped from 6326.991 to 2494.657 because of 1 cluster increased.

Furthermore, the decreasing value is followed normally which indicates there is no increasing performance anymore. Finally, an *elbow method* was applied to determine the number of clusters. The Elbow method draws a variety range of clusters (K) that resembles an arm. Then, the point which reflects an elbow indicates the near-optimum number of clusters. Therefore, based on this result, we decide to generate only 3 clusters, the same number of clusters developed by the X-means algorithm. To have more insight, we visualized the result in figure 4.

Table 5. Average within Cluster Distance

Number of Clusters	Average within-cluster distance
2	-6326.991
3	-2494.657
4	-1664.941
5	-957.870
6	-740.551
7	-628.412

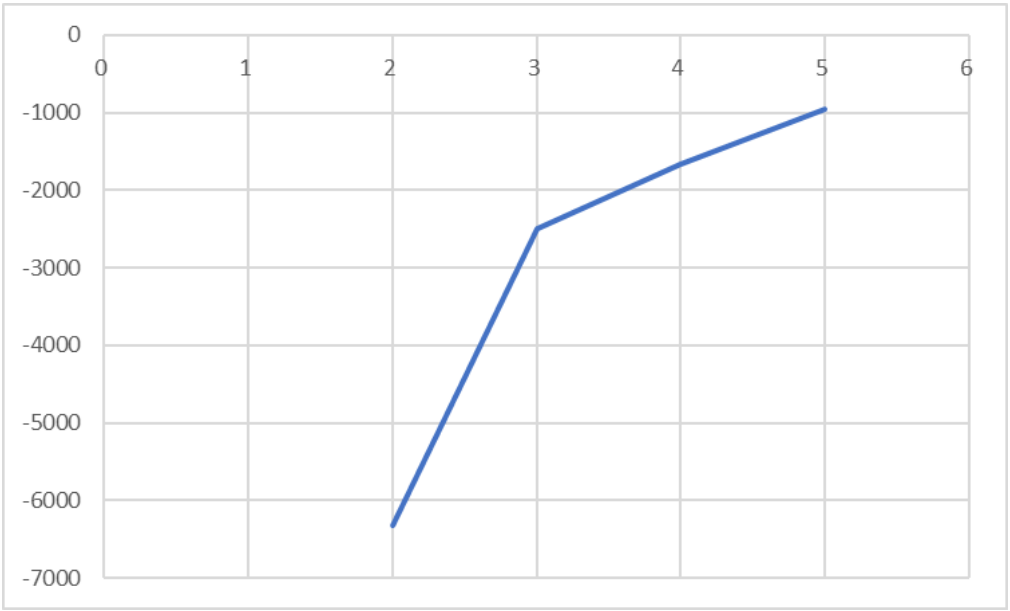

Figure 4. Average within-cluster distance

The results of real estate segmentations are described in table 6. From table 6, we can see that within cluster 0, there are 15.83%, 85.77%, 56.66% of high, low, and medium price levels, respectively. It means, low price dominates cluster 0. Within cluster 1, there are 2.08%, 1.17%, 2.45% of high, low, and medium price levels, respectively which is quite balance compared to cluster 0. In cluster 2, high, low and medium cluster are 82.08%, 13.05%, 40.89% respectively. Therefore, in cluster 2, high price dominates cluster 2. Based on this result, we can conclude that, in terms of house prices, cluster 0 is characterized by a low house price cluster and cluster 2 is characterized by high house prices. While cluster 1 is categorized as uncommon house prices (outlier), only a few data houses are occupied. The result of house clustering is also visualized in figure 5.

Table 6. Clustering Results

Price level	Cluster	Count	Percentage
High	cluster_0	38	15.83%
Low	cluster_0	657	85.77%
Medium	cluster_0	532	56.66%
High	cluster_1	5	2.08%
Low	cluster_1	9	1.17%
Medium	cluster_1	23	2.45%
High	cluster_2	197	82.08%
Low	cluster_2	100	13.05%
Medium	cluster_2	384	40.89%

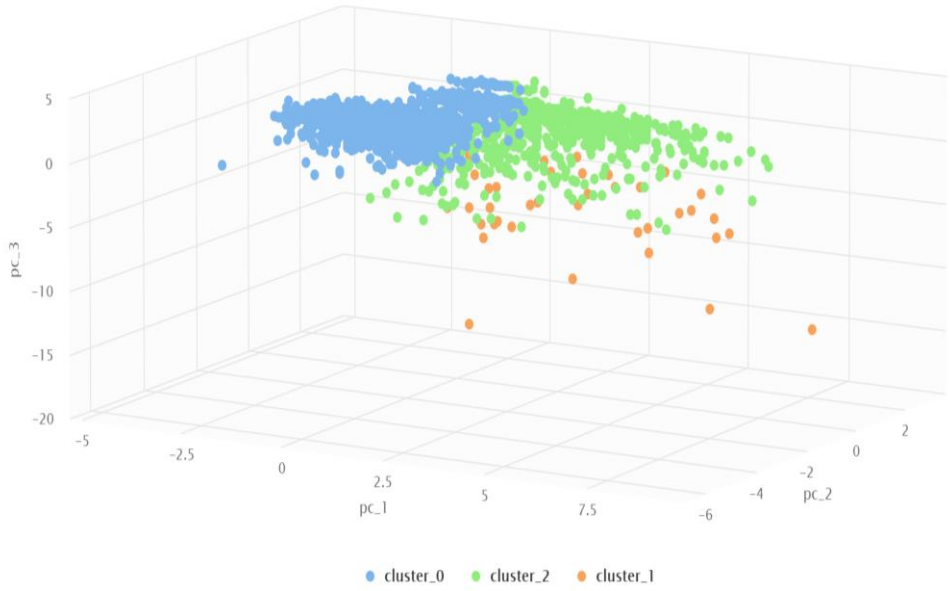

Figure 5. Houses Clustering Result

To have more information regarding the characteristic of clusters 0, 1, and 2, we run a classification task for predicting the house price ranges (low, medium, and high) and weigh the variable importance of sold house characteristics for determining the most important characteristic in cluster 0,1 and 2. For this purpose, we employed a random forest algorithm to analyze the characteristic of each cluster. However, several parameters need to be tuned therefore we performed a grid search technique in achieving near optimum parameters.

4.1. Cluster 0

To have the attribute importance of cluster 0, we performed a random forest algorithm combined with a grid search technique. A grid search technique was applied in searching near optimum parameters. The used parameters here are the number of trees and criterion. The number of trees ranges from 11, 30, 50, 70, and 100. Moreover, we also examined 3 criteria: information gain, information gain ratio, and Gini index. The results of parameter optimizations are described in table

7. The model achieves 82% accuracy by using 50 trees and information gain as a criterion.

Table 7. The result of grid search for cluster 0

Number of tress	Criterion	Accuracy
50.0	information_gain	0.8209677419354838
100.0	gain_ratio	0.8193548387096774
30.0	gain_ratio	0.8153225806451614
70.0	gain_ratio	0.810483870967742
70.0	gini_index	0.810483870967742
70.0	information_gain	0.807258064516129
100.0	information_gain	0.807258064516129
50.0	gain_ratio	0.8056451612903224
100.0	gini_index	0.8024193548387097
50.0	gini_index	0.7951612903225806
30.0	information_gain	0.7911290322580644
30.0	gini_index	0.7887096774193547
11.0	gain_ratio	0.7822580645161289
11.0	information_gain	0.7798387096774194
11.0	gini_index	0.7790322580645161

With 82% of prediction accuracy, the model was then used to weigh the attribute importance. For example, in predicting the house prices in cluster 0, latitude coordinate is found as the most important attribute followed by the square footage of the house, square footage of the lot, and longitude coordinate in rank 2,3, and 4 respectively. Therefore, we can conclude that cluster 0 which is dominated by the low price level (segmentation result) is characterized by the latitude coordinate, square footage of the house, square footage of the lot, and longitude coordinate. The result of attribute weighting is shown in figure 6.

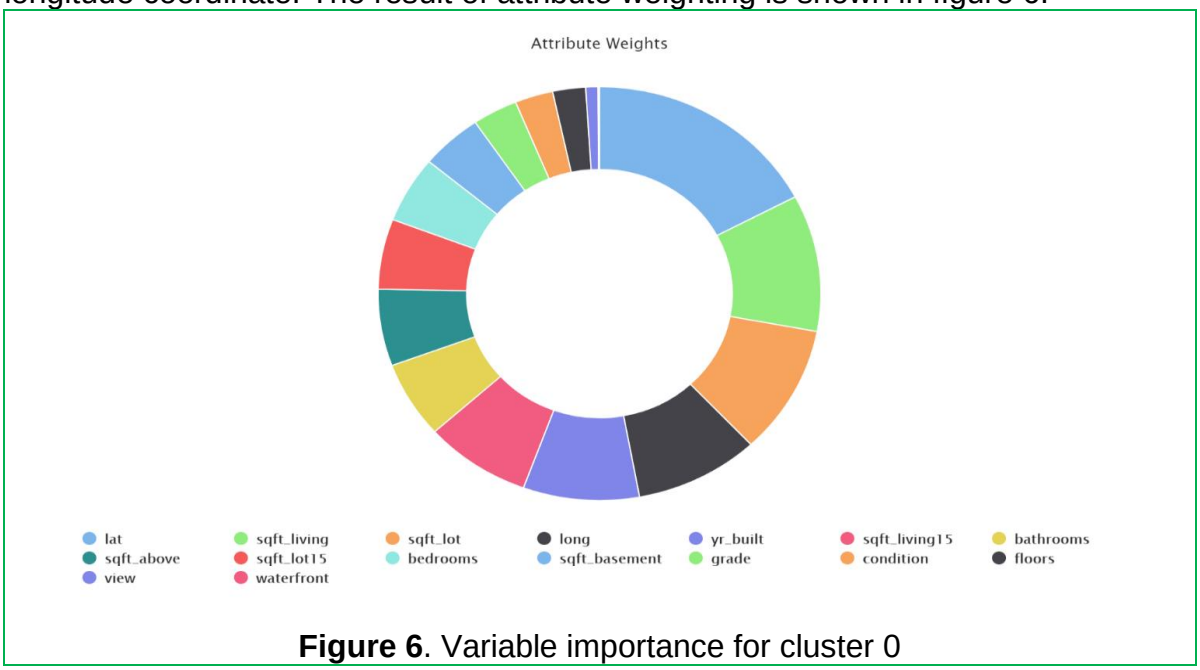

Figure 6. Variable importance for cluster 0

4.2. Cluster 1

Based on the result of k-means clustering, cluster 1 was found as an uncommon cluster. Besides, this cluster also consists of 37 houses considered the least cluster compared to cluster 0 and cluster 1. Therefore, to gain more insight into this cluster, we performed a classification task using a random forest algorithm. We also employed a grid search technique to optimize some parameters. The result of the grid search is described in table 7. Eventually, leveraging 11 tresses and using the Gini index as a criterion, the model achieves 79% accuracy. As a result, stated in table 8, more trees do not guarantee more accuracy. This may happen because of the unnormal data or outliers in cluster 1.

Table 8. The result of grid search for cluster 1

Number of trees	Criterion	Accuracy
11.0	gini_index	0.7916666666666667
50.0	gini_index	0.775
11.0	information_gain	0.7583333333333333
100.0	gain_ratio	0.7416666666666667
30.0	information_gain	0.7416666666666667
50.0	information_gain	0.7416666666666667
100.0	information_gain	0.7333333333333334
30.0	gain_ratio	0.7333333333333333
100.0	gini_index	0.725
11.0	gain_ratio	0.7166666666666667
70.0	information_gain	0.7166666666666667
70.0	gini_index	0.7083333333333334
30.0	gini_index	0.7083333333333333
50.0	gain_ratio	0.6833333333333333
70.0	gain_ratio	0.6833333333333333

After finishing the house price prediction, we weighted the attribute importance in cluster 1. As the result stated before, cluster 1 contains uncommon behavior which also results in uncommon cluster characteristics. For example, the number of bathrooms is found as the most important attribute in predicting house prices followed by the square footage of the house, square footage of the lot, and the number of bedrooms with rank 2,3 and 4 respectively. As an uncommon cluster, this result also considers bright future research. The result of attributes importance is described in figure 7.

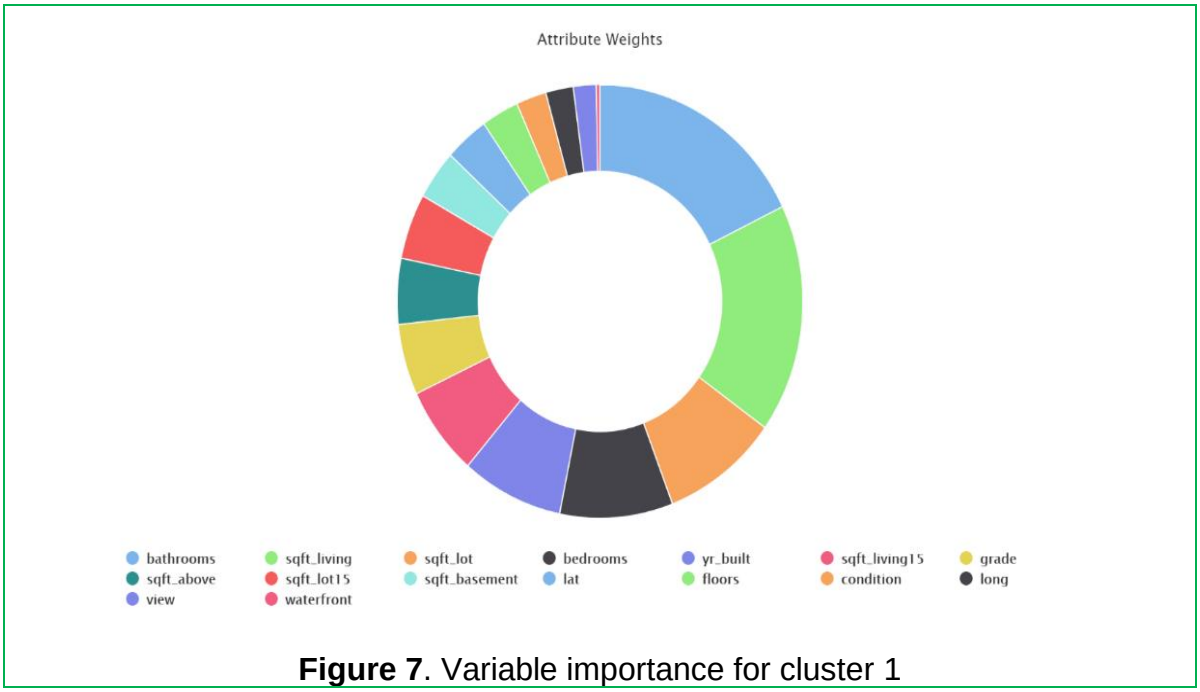

Figure 7. Variable importance for cluster 1

4. 3. Cluster 2

The same treatment was applied to house data in cluster 2. To predict the house prices, a random forest algorithm was utilized, and to have near optimum parameters, a grid search technique was employed. From 3 types of random forest criteria, information was found as the most accurate criterion combined with 70 trees used in the system. Eventually, the model achieves 81.3% accuracy which is slightly better than cluster 1 accuracy. To have more understanding regarding cluster 2, we then performed an attribute weighting analysis. The complete results of the grid search for cluster 2 were drawn in Table 9.

Table 9. The result of grid search for cluster 2

Number of trees	Criterion	Accuracy
70.0	information_gain	0.8131840796019901
30.0	gini_index	0.8129579375848033
70.0	gini_index	0.8116915422885572
100.0	gini_index	0.8116689280868385
30.0	information_gain	0.8013116236996833
100.0	information_gain	0.7982134780642242
11.0	gini_index	0.7952962460425148
50.0	gain_ratio	0.7952962460425146
50.0	gini_index	0.7951153324287652
50.0	information_gain	0.7936906377204884
11.0	information_gain	0.7847580280416102
70.0	gain_ratio	0.7847127996381728
100.0	gain_ratio	0.7817729534147445
30.0	gain_ratio	0.7773405698778832
11.0	gain_ratio	0.768340117593849

The result of attribute weighting is drawn in figure 8. In cluster 2, latitude coordinate was found as the most important characteristic then followed by the square footage of the house, square footage of the lot, and square footage of the house apart from the basement with rank 2, 3, and 4 respectively. From the results, it is confirmed that each cluster has its own characteristic which then can be utilized by the appraiser or even broker and agent for marketing purposes. Those sold houses segmentation also reflect the preference of customers with respect to which cluster they are.

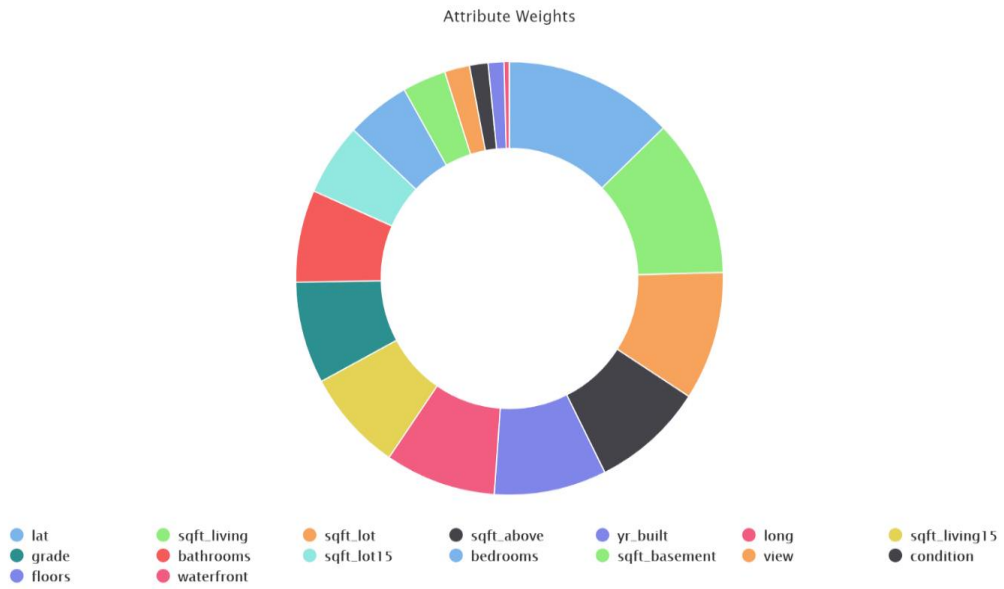

Figure 8. Variable importance for cluster 2

5. Managerial Implication

As a managerial implication, a map for the decision support system was drawn and shown in figure 9. This process started by feeding the data into the system. At this stage, the data were preprocessed. Afterward, three potential clustering algorithms are examined and compared in terms of processing time and average within-cluster distance. A robust clustering algorithm is then used to cluster the data. Eventually, three clusters are found and to understand the characteristic of each cluster, a classification task was performed for each cluster by using a random forest algorithm.

Refer to Silver (Silver, 1991), there are two forms of guidance in a decision support system intended to help decision-makers such as suggestive and informative guidance. Hence, we utilize the variables weighing results for helping decision-makers achieve effectiveness. Suggestive guidance consists of several main points such as selecting and designing the recommended and not recommended value, in this case, the most essential characteristics of houses for customers with respect to their clusters are considered. On the other hand, for the informative guidance purpose, an appraiser, broker, and agent enjoy important information such as the description of the required input value to get the houses attracted to customers, the description on how to use the input, and the similar behavior of customers, etc. Therefore, applying this decision support system is

crucial in helping appraisers, brokers, and agents promote the houses to potential consumers.

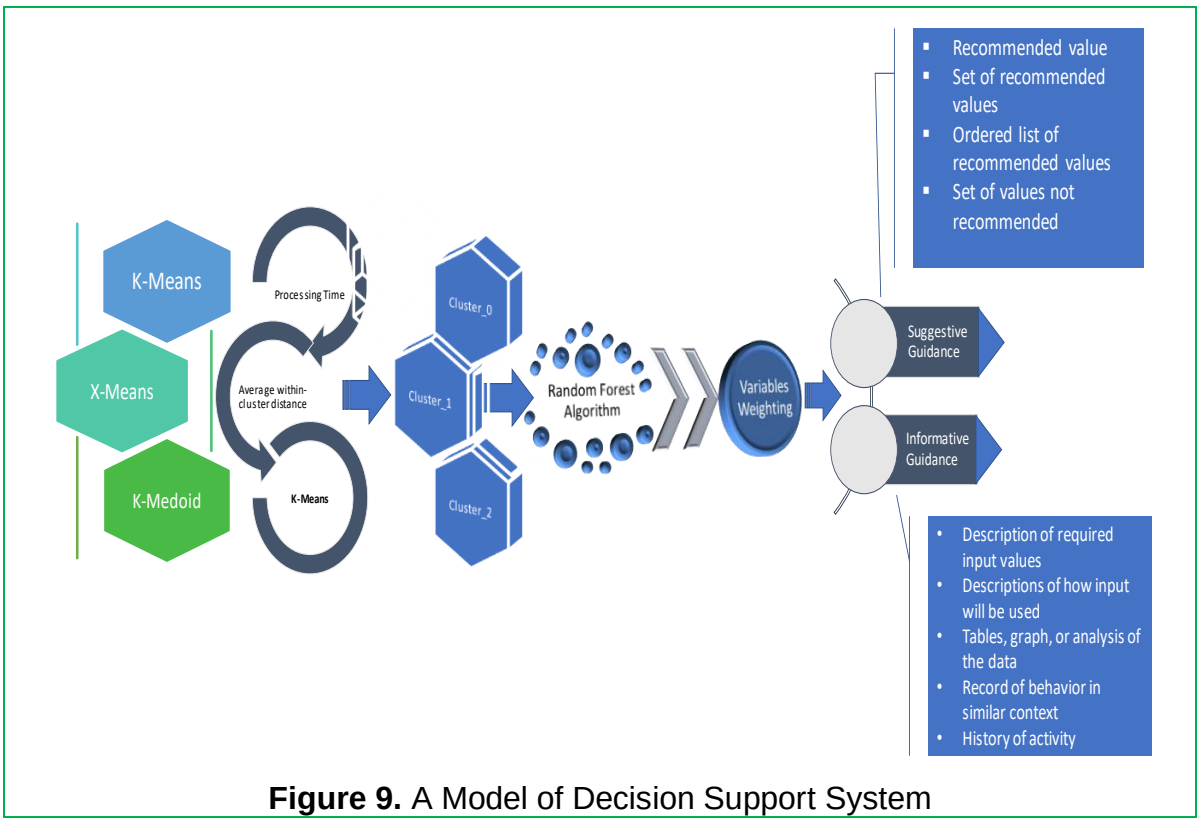

Figure 9. A Model of Decision Support System

6. Conclusion

Appraisers, brokers, and agents of real estate often fall into subjective decision-making. Therefore, more complete information as a decision support system is urgently needed. We assessed and compared K-means, X-means and K-medoid algorithm performances for clustering 18 sold houses characteristics using 21,613 observations of sold houses to tackle this problem. Processing time and average within-cluster distance are used as a measurement scale. To reduce the model dimensionality, we performed a Principal Component Analysis which then result in 3 principal components. Afterward, these principle components are clustered using the k-means, k-medoid, and x-means algorithm. Eventually, K-means and X-means reveal a very close performance on processing time and Average within-cluster distance with 14s processing time, and 2494.657, 2499.439 average within-cluster distance respectively. However, K-medoid requires more processing time (73s) and a larger Average within-cluster distance (6341.951). Therefore, we can conclude that K-means is slightly better than the X-means algorithm.

To have more complete information regarding the characteristic of each cluster, a classification task which respects to house prices was conducted. We employed a random forest algorithm, to predict the house prices. In order to avoid the overfitting problem, we applied the 10-fold cross-validation technique combined with the grid search algorithm for having near optimum parameters. We also optimized the number of trees in the forest (11, 30, 50, 70, and 100 trees) and the

types of criterion (information gain, information gain ratio, and Gini index) using the grid search technique.

Using the K-means algorithm, 3 clusters were identified: cluster 0, cluster 1, and cluster 2. In cluster 0, low house prices dominate the cluster, with 85.77% of total low house prices. Within this cluster, the 3 most important characteristics are lat (latitude coordinate), square ft living, square ft lot, and longitude coordinate which practically meant latitude coordinate factor is considered as the most important factor for consumers in low house price segmentation. However, cluster 1 previously considered an uncommon cluster is not dominated by any certain consumers in the house price level. This cluster is characterized by the 3 most important attributes: the number of bathrooms, square ft living, and square ft lot, which are considered an outlier or uncommon condition. In cluster 2, more high house prices are clustered here with 82.08% of total high house prices. This cluster is characterized by lat (latitude coordinate), square ft living, square ft lot, and square ft above. These characteristics are very similar to the characteristic of cluster 0, the only difference is the fourth most important attribute. The complete results are described in table 10.

Table 10. Results of Clustering and Its Characteristic Importance

No	Price level	Cluster	Count	Percentage	Important Characteristics
1	High	cluster_0	38	15.83%	1. Lat 2. Square ft living 3. Square ft lot 4. Longitude coordinate
	Low	cluster_0	657	85.77%	
	Medium	cluster_0	532	56.66%	
2	High	cluster_1	5	2.08%	1. Bathrooms 2. Square ft living 3. Square ft lot 4. Bedrooms
	Low	cluster_1	9	1.17%	
	Medium	cluster_1	23	2.45%	
3	High	cluster_2	197	82.08%	1. Lat 2. Square ft Living 3. Square ft Lot 4. Square ft above
	Low	cluster_2	100	13.05%	
	Medium	cluster_2	384	40.89%	

The phenomena of uncommon and outliers of cluster 1 could be potential future research. Outliers detection and behavioral understanding are two very interesting topics in the real estate field. Moreover, the recent advanced development of information and technology allows researchers and practitioners to have more supplies in pursuing more accurate and reliable research results. Results of Clustering and Its Characteristic Importance are described in table 10.

References

Abbasimehr, H., & Shabani, M. (2019). A new methodology for customer behavior analysis using time series clustering: A case study on a bank's customers. *Kybernetes*.

Antipov, E. A., & Pokryshevskaya, E. B. (2012). Mass appraisal of residential apartments: An application of Random forest for valuation and a CART-based approach for model diagnostics. *Expert Systems with Applications*, *39*(2), 1772–1778.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Brusco, M. J., Singh, R., Cradit, J. D., & Steinley, D. (2017). Cluster analysis in empirical OM research: survey and recommendations. *International Journal of Operations & Production Management*.

Burt, M., & Liew, C. L. (2012). Searching with clustering: An investigation into the effects on users' search experience and satisfaction. *Online Information Review*.

Chhetri, P., Han, J. H., & Corcoran, J. (2009). Modelling spatial fragmentation of the Brisbane housing market. *Urban Policy and Research*, *27*(1), 73–89.

de Aquino Afonso, B. K., Melo, L. C., de Oliveira, W. D. G., da Silva Sousa, S. B., & Berton, L. (2020). Housing prices prediction with a deep learning and random forest ensemble. *Unpublished Manuscript]. Anais Do Encontro Nacional de Inteligencia Artificial e Computacion*.

el Moudden, I., Ouzir, M., Benyacoub, B., & ElBernoussi, S. (2016). Mining human activity using dimensionality reduction and pattern recognition. *Contemporary Engineering Sciences*, *9*(21), 1031–1041.

Esmaeilzadeh, M., Abdollahi, B., Ganjali, A., & Hasanpoor, A. (2016). Evaluation of employee profiles using a hybrid clustering and optimization model: practical study. *International Journal of Intelligent Computing and Cybernetics*.

Fourati-Jamoussi, F., Niamba, C.-N., & Duquennoy, J. (2018). An evaluation of competitive and technological intelligence tools: A cluster analysis of users' perceptions. *Journal of Intelligence Studies in Business*, *8*(1).

Geltner, D. (1989). Bias in appraisal-based returns. *Real Estate Economics*, *17*(3), 338–352.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

Hoesli, M., Lizieri, C., & MacGregor, B. (1997). The spatial dimensions of the investment performance of UK commercial property. *Urban Studies*, *34*(9), 1475–1494.

Kang, J. M. (1995). A Study on Market Segmentation of Apartment in Seoul. *Master D. Diss. University of KonKuk*.

Kaufman, L., & Rousseeuw, P. J. (2009). *Finding groups in data: an introduction to cluster analysis* (Vol. 344). John Wiley & Sons.

Kim, B. T. (2000). Determination of spatial boundaries of Seoul housing sub-market. *Master D. Diss. University of YonSei*.

Kim, G. S., & Park, J. Y. (2003). The spatial pattern of housing prices: Seoul and new towns. *Journal of the Korean Regional Science Association*, *19*, 47–61.

Kuşan, H., Aytekin, O., & Özdemir, İ. (2010). The use of fuzzy logic in predicting house selling price. *Expert Systems with Applications*, *37*(3), 1808–1813.

Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, *28*(2), 129–137.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, *1*(14), 281–297.

Meen, T.-H., Prior, S. D., Lam, A. D. K.-T., Zhu, M., Wang, W., & Huang, J. (2014). Improved initial cluster center selection in K-means clustering. *Engineering Computations*.

Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *2*(11), 559–572.

Pelleg, D., & Moore, A. W. (2000). X-means: Extending k-means with efficient estimation of the number of clusters. *Icml*, *1*, 727–734.

Peterson, S., & Flanagan, A. (2009). Neural network hedonic pricing models in mass real estate appraisal. *Journal of Real Estate Research*, *31*(2), 147–164.

Selim, H. (2009). Determinants of house prices in Turkey: Hedonic regression versus artificial neural network. *Expert Systems with Applications*, *36*(2), 2843–2852.

Silver, M. S. (1991). Decisional guidance for computer-based decision support. *MIS Quarterly*, 105–122.

Simon, H. A. (1957). *Models of man: social and rational; mathematical essays on rational human behavior in society setting*. New York: Wiley.

Tidwell, O. A. (2011). *An investigation into appraisal bias: The role of decision support tools in debiasing valuation judgments*.

Vasan, K. K., & Surendiran, B. (2016). Dimensionality reduction using principal component analysis for network intrusion detection. *Perspectives in Science*, *8*, 510–512.

Wu, C., & Sharma, R. (2012). Housing submarket classification: The role of spatial contiguity. *Applied Geography*, *32*(2), 746–756.

Wu, C., Ye, X., Ren, F., & Du, Q. (2018). Modified data-driven framework for housing market segmentation. *Journal of Urban Planning and Development*, *144*(4), 4018036.

Wu, L., & Brynjolfsson, E. (2015). The future of prediction: How Google searches foreshadow housing prices and sales. In *Economic analysis of the digital economy* (pp. 89–118). University of Chicago Press.

Wu, Q., Zhan, C., Wang, F. L., Wang, S., & Tang, Z. (2016). Clustering of online learning resources via minimum spanning tree. *Asian Association of Open Universities Journal*.

Yatchev, I., Stancheva, R., Marinova, I., Jaafar, A., Sareni, B., & Roboam, X. (2012). *Clustering analysis of railway driving missions with niching. COMPEL-The international journal for computation and mathematics in electrical and electronic engineering*.

Zhong, X., & Enke, D. (2017). Forecasting daily stock market return using dimensionality reduction. *Expert Systems with Applications*, *67*, 126–139.