Data Mining Application Forecast of Business Trends of Electronic Products

Kheo Chau Mui, Nhon Nguyen Thien
Information Technology Department, FPT University, Cantho city, Vietnam

***Abstract*—Sales forecasting is a pressing concern for companies amid rising consumer demand and intensifying competition, compounded by declining sales due to growing socio-economic challenges. Currently, many companies are having difficulty selling products due to a lack of management systems. To assist that, data mining techniques are introduced but it is difficult to evaluate the data and it is practically impossible to accurately forecast large amounts of data. However, data mining remains an important management tool that supports early decisions to increase profits, innovate business trends and improve sales by generating intelligence from the company's data resources. In this article, the research object chosen is the data of a nationwide electronics company. Their sales volume data for consumer electronics was used and applied to this study. The study used a "clustering" algorithm to group data based on the unique characteristics of each product, region, season, and time to estimate the amount of goods sold in the past, thereby predicting the amount of goods that will be exported. Password is sold in the following years and look for market trends. For each group, the results obtained with k = 3 show that the number of elements in each cluster is 771422, 11874, and 312, respectively. Combined with the "regression tree" algorithm for cluster partitioning and using the protocol Evaluate MSE and RMSE to evaluate the accuracy of the model, a result of 43065.66 Sales forecasting results show that the model's accuracy is close to realistic accuracy and depends on seasonal factors that are really important to some people. Based on the above results, the business's marketing campaigns and strategies will be deployed and achieve high results.**

Keywords—Data mining; sales forecasting; clusters; regression tree; RMSE; MSE; k-prototypes

I. Introduction

Electronic gadgets have steadily become important goods to assist people's life in the contemporary era of smart technology. This generates a lucrative profit for firms dealing in electronic items; as rivalry between businesses grows, they must always come up with new business advantages to compete [40], [41]. In the electronics business, you must compete and survive. These benefits must be based on the quantity of data gathered in the past and present from internal operations, product supply procedures, market trends and the business environment, and consumer preferences, to analyse patterns, estimate future sales, and establish a change management approach that brings efficiency and quality to corporate management while saving operational costs and expenses. Costs of storing are reduced, and profits are boosted [1], [5]. However, achieving high company efficiency through data mining, analysis, and trend prediction is a challenging task for firms. As a result, a system is required for firms to efficiently explore, analyse, process, and anticipate new business trends [27], [28].

Previous research has offered an overview of contemporary obstacles in sales forecasting as well as difficulties in trading electronic products, such as:

- Product life cycle is becoming more shorter.
- Consumer demand has increased and become more diverse.
- The average industrial land rental price rises by 5-8% every year.

To be very lucrative, a corporation must properly estimate the output of commodities, at the right moment of demand, and restrict inventory. Businesses will have challenges without the tools of information processing and analysis to assist anticipate the next business condition if they have a large and diversified data source. It is important to develop a "Data mining application to forecast business trends for electronic products" based on relevant research and existing business practices. This programme will assist businesses in selecting a strategy to anticipate the optimal output of items for their firm based on each area, industry, and seasonal features.

II. Related Work

For many years, several methodologies have been utilised in sales forecasting research. Here are some examples of typical studies. Pure Classification (PC) and Hybrid Clustering Classification (HCC) are two data mining techniques suggested by Bhavin Parikh et al. [4]. The results of the tests demonstrate that the HCC model outperforms the PC model in terms of accuracy and performance. In particular, after evaluating 500 samples with Nearest Neighbours = 5, the accuracy attained is 57.62%, outperforming the PC model. Fifit Alfiah et al. [2] investigated association rules in data mining with excellent dependability. Support: 0.1 and support x Confidence: 0.05 are the results, and if the manager simply enters the forecast quantity of 13 goods, the output forecast quantity is support: 0.2, support x confidence: 0.1, and prediction percentage: 0.15. The quantity value is enhanced by 15% from the initial value based on the outcomes and association rules in data mining. A paper in [3] uses the K-means algorithm to divide data into three separate clusters based on product type and sold quantity, namely Dead-Stock (inventory), Slow-Move (sold products), and rapid-Move (rapid sale). Next, employ the MFP: Most Frequent Pattern algorithm to identify frequent item attribute patterns in each product category while also providing sales trends in a concise

format. Lytvynenko Tetiana's [6] research employed two key methodologies: statistical and structural methods, respectively, along with the widely used Decision Tree method to represent the process visually and simply on java. Simplify the main goal of the analysis. They utilised a data set that contained the sales volume of an employee job, the month of sales, and the sales of a business from 2006 to 2009. After sifting through the binary tree, they discovered that April, November, and December had the largest sales (about 23000 units each month), although accounting for 18771 in 2006. 19139 and 15164. Mustapha Ismail et al. [5] conducted research on data mining (DM) for e-commerce, which included three general algorithms: matching, grouping, and prediction. It also discusses some of the advantages of DM for e-commerce businesses, including as item planning, sales forecasting, shopping cart analysis, customer relationship management, and probable market segmentation obtained by the three techniques listed above. Furthermore, this research assesses data mining difficulties such as spider detection, data translation, and making the data model intelligible to business users.

III. Related Methods

To implement the application, we choose to utilise the "K-means" and "K-Prototypes" algorithms to cluster data based on the characteristics of each product, region, and time, in conjunction with the "Regression tree" technique. From [7], [8], [9], [12], [19], [20], [21], [24], [25], [26]. Python is the programming language used in this application. This combination aids in forecasting development patterns, product sales to address the issue of shortages, not keeping up with the seasons, consumer purchasing habits, and undesirable inventory that firms frequently face. Management and operation must entail significant expenditures.

A. K-means

J.A. Hartigan and M. A. Wong of Yale University created the Kmeans algorithm [18]. It is one of the finest algorithms for forming clusters of tiny values from vast amounts of unlabeled data (Label). The Kmeans method is an iterative algorithm that attempts to locate data clusters that are as close as feasible.

The initialization of the number of k groups is the first stage in the Kmeans algorithm [13], [14], [15], [16], [17]. And the starting centre value for each group is chosen at random.

The distance between each element and the centre values of each group is then calculated. Different formulae will be employed to calculate the distance depending on the properties of each data type. Manhattan, Cosine, Minkowski, and Euclidean distance metrics are utilised for numerical data. The Euclidean distance is employed in this research to compute the distance from the centre to the elements:

$$\boldsymbol{d(x_i, x_j)} = \sqrt{\left(|\boldsymbol{x_{i1}} - \boldsymbol{x_{j1}}|^2 + |\boldsymbol{x_{i2}} - \boldsymbol{x_{j2}}|^2 + \cdots + |\boldsymbol{x_{in}} - \boldsymbol{x_{jn}}|^2\right)}$$

In n-dimensional space, xi=(xi1,xi2,...,xin) and xj=(xj1,xj2,...,xjn) are two data points.

Then, until convergence, perform the following operations:

+ Based on distance, each element is assigned to its nearest centre.

+ Update the centres of K clusters, each centre being the mean value of the elements in its cluster.

B. K-prototypes

Separates the dissimilarity of the combined data into two sections for independent computations. The numerical component employs squared Euclidean distance, whereas the mixed part uses basic mactching distance [37, 38, 39] Because the ratio of the two data types is not the same, the study will alter the parameters in the K-Prototypes method after the calculation to avoid the divergence of the grouping result value. The distance is defined as follows, where m is the number of matches and p is the total number of variables (attributes):

$$d(i,j) = \frac{p-m}{p}$$

The K-Prototypes algorithm is implemented as follows:

Input: Initial data set X and number of clusters k.

Output: k sample objects so that the standard function approaches the minimal value.

1) Create k initial sample objects for X, each of which serves as the representative centre of each cluster.

2) Distribute each X feature to each cluster so that it is closest to the sample object in the cluster, while updating the sample object for each cluster.

3) After all of the objects have been dispersed to the clusters, compare their similarity to the sample objects to see if there is a sample object most similar to it that varies from the other. The current cluster's sample object moves the object under consideration to the cluster corresponding to the sample object nearest to it and simultaneously changes the sample objects for these two clusters.

4) After inspecting all objects, repeat step 3 until no object changes.

C. Elbow Method

The Elbow technique [22] is one approach for determining the ideal number of clusters k while clustering. This approach is based on Thorndike's [23] hypothesis, and Elbow is a visual method. The SSE (Sum of Squared Error) indicator represents this Elbow technique.

The basic idea behind this approach is to compute k values by squaring the distances between the members in each cluster and the cluster centre. The sum of squared errors (SSE) is used for comparison. Repeat the k value and compute the SSE; smaller values suggest that each cluster is more convergent.

$$SSE = \sum_{k=1}^{K} \sum_{x_i \in S_k} ||x_i - C_k||_2^2$$

where: k is the number of clusters, Ck is the kth cluster, and x is the number of cluster elements.

The Elbow approach may be defined in two ways:

1) Deformation: It is determined as the mean of the squared distances from the individual cluster centres. The Euclidean distance measure is commonly employed.

2) Inertia: The sum of the squares of the sample distances from the nearest cluster centre.

Iterate over the values of k from 1 to 9 and compute the distortion and inertia for each value of k in the specified range shown in Fig. 1.

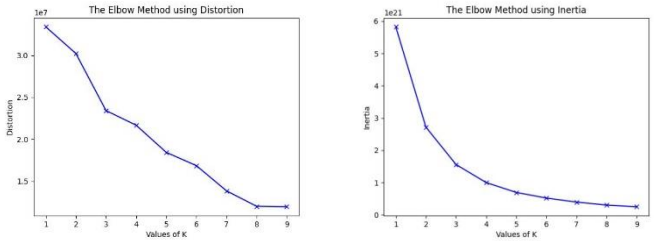

Fig. 1. An illustration of the elbow method.

D. Regression Tree Method

The regression tree approach is used to forecast continuous label values like revenue, profit, product cost, and so on.

An algorithm-based regression tree is used to estimate sales trends ([10]). In terms of current advancements in regression tree approaches, see [29], [30], [31], [32], [34], [35], [36.Regression trees are built using a method known as binary recursive partitioning, which is an iterative procedure that divides data into branches and then further sub-branches. Because of its simplicity, the regression tree technique published by (Breiman et al., 1984; Quinlan, 1993) is an automated machine learning model that is frequently used in data mining (Wu and Kumar, 2009). Calculate the standard deviation (Standard Deviation) and assess the dispersion of a data set using the regression tree technique. A big standard deviation indicates that the data has a high degree of dispersion and variability.

The symbol for Standard Deviation is σ (Sigma)

And then compute the standard deviation by taking the square root of the variance.

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

The standard deviation calculation procedure consists of four steps:

1) Standard.
2) Average and squared results for each number
3) Next, compute the mean of the squared differences.
4) Multiply all values by the square root.

After calculating the standard deviation, we use the regression tree approach to split the data.

The regression tree algorithm goes through the following steps: [11]

1) Begin with a single node that contains all of the points. (Calculate standard deviation using the formula)

2) Stop if all node points have the same value for all independent variables. Searching for a variable over all binary divisions of all variables, on the other hand, will lower Sigma as much as feasible.

3) Repeat step 1 for each new node.

We proceed to analyse the model when we have obtained the findings.

E. Equation Methods

The RMSE, MSE, and MAE indices are often used to evaluate the accuracy of a regression issue. The standard deviation of the residuals (the prediction error) is the Root Mean Square Error (RMSE). The residual is a measure of how far the data points are from the regression line; the RMSE is a measure of how diffuse these residuals are. In other words, it indicates how dense the data is around the best-fit line. In climate research, forecasting, and regression, the base mean square error is frequently used to validate experimental results. The RMSE assessment technique is utilised in this work according to the formula [33]:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(p_i - r_i)^2}$$

The independent variable is pi, and the predicted value is ri.

When the system is performing optimally, these measurement values approach zero. The greater this value, the poorer the system's efficiency.

IV. Methodology

A. Datasets

The data represents the sales results for the three years 2017, 2018, and 2019 from a company that specializes in selling electrical and electronic equipment on a national basis. With almost 10 million goods distributed annually. Includes product groups: Home entertainment equipment; Household products; Kitchen products; Air conditioning; health and beauty support equipment, etc. Predicted data is the result of the number of goods sold each year. Table I shows some information of the data.

TABLE I. Some Information About Industry Codes in 2018

No	Product code (explanation)		Quantity
1	BEAUTY	Beauty equipment	36
2	C-BATT	The battery	60
3	COLDCHAIN	Freezer	46
…	……	……	…
26	TELEPHONE	Phone	34
27	VC	Vacuum cleaner	17
28	WM	Washing machine	56

B. Algorithm Diagram

The research model consists of two stages in Fig. 2: Stage 1: using the "clustering" algorithm to cluster data according to 4 main characteristics: product type, time, sales quantity and product consumption location. Phase 2: use the prediction method using the "Regression tree" algorithm to build a system to predict product development trends/sales.

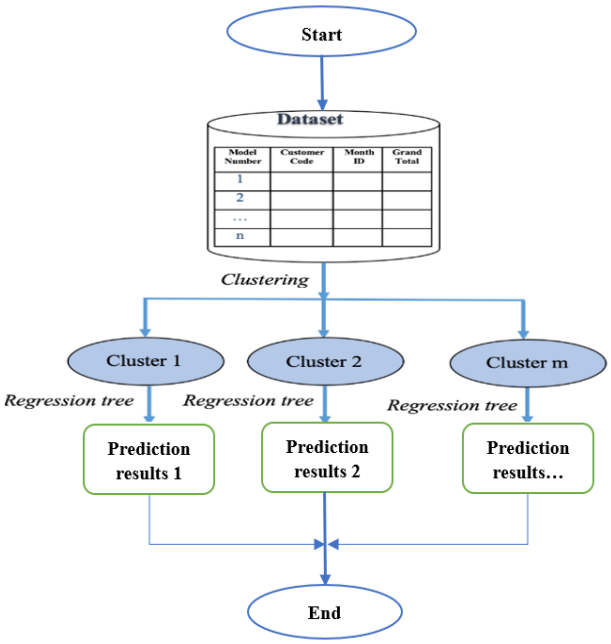

Fig. 2. Algorithm diagram.

C. Data Preprocessing

In practise, data is frequently heterogeneous in terms of data type, with data consisting of several properties (columns) with varying units and magnitudes. This has an impact on the efficiency of many algorithms, as well as their correctness. As a result, before used, the data must be normalised. When the data domains of mixed attributes differ substantially, normalising LabelEncoder data from the sklearn package is a technique widely used as part of preprocessing. The purpose of normalisation is to convert the values of discrete text-valued columns in the data set to a common scale while preserving the range of values. Because of the discrete nature of the data, it is necessary to normalise the values of attributes such as commodity code (Model Type), customer code (Customer Code), month code (Month ID), and total number of items sold (Grand Total) on a common scale in order to generate predictions. Fig. 3 shows the Numeric Data type after standardization, and Fig. 4 is represented as a quantity distribution chart by product type.

D. Clustering (State 1)

With a data set of goods that includes 30 product groups and four different attributes such as commodity code (Model Type), customer code (Customer Code), month code (Month ID), and total number of products sold (Grand Total), the algorithm creates the centre value or calculates the distance in turn before moving on to the next group.

customerCode	LabelEncode
25103	0
25231	1
270112	2
27516	3
27535	4
27597	5
39140	6
5000003464	7
5000003469	8
5000003470	9
5000003471	10

modelType	LabelEncode
C-BATT	0
COLDCHAIN	1
COMAC	2
COMPACT	3
DSC ACC	4
DSC DSLM	5
E-IRON	6
E-SHOWER	7
HD CAM	8
HEADPHONE	9
HEALTHCRE	10

Fig. 3. Data after normalization in numeric form.

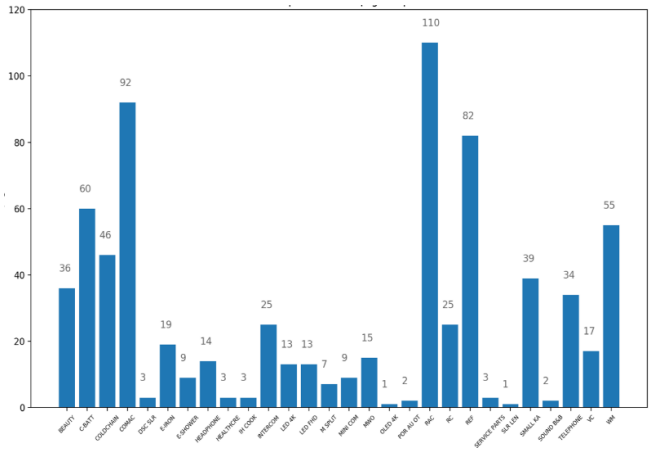

Fig. 4. Quantity distribution chart by product type in 2018.

The clustering algorithm's initial step is to employ k-means [10, 22, and 15]. We set the initial centre values for the groupings. If the data is clustered into k clusters, the kmeans_init_centers function chooses a starting point at random from the data set (Model Number, Customer Code, Month ID, Grand Total).

Specifically, dA(i,j) represents the distance based on the Month ID and Total Grand values, whereas dC represents the distance based on the Model Number and Customer Code characteristics. d(i,j) is determined using the Euclidean distance and two random numbers i and j.

$$\mathrm{dA(i,j)} = \sqrt{(|x_i - x_j|)^2}$$

Because the K-means method only works on numeric data and the remaining characteristics have mixed data types, the K-Prototypes algorithm should be used as shown below.

$$dC(i,j) = \frac{p-m}{p}$$

where, m is the total number of variables (attributes) and p is the number of matches

After determining the two distances dA and dC, compute the common distance for two values of i and j as follows.

$$d(i,j) = m * dA(i,j) + (1-m) * dC(i,j)$$

This formula will return values that are comparable. The distribution of items in the group is extremely skewed in certain circumstances where there are no elements in the

group. This problem is solved by repeating the clustering procedure until no empty groups remain.

E. Predict (State 2)

Building a system to anticipate the development trend / product sales using the "Regression tree" algorithm's prediction approach.

After clustering, the findings are discovered to be groupings of cluster data aspects of products, with each cluster having unique qualities. The user will choose which cluster he or she belongs to before applying the Regression Tree issue to each cluster to compute the likelihood of correctly forecasting the quantity of goods sold [10].

Algorithm: Regression tree
Input: - Dataset: data set S each cluster (cluster 1, cluster 2, cluster 3)
Output: - Result tree of each cluster
- S = {s_1, s_2, …s_k} is a set Sigma. - Initialize the set of all points: Sk: f(x) = {x1, x2, …xn}. xi are the points in the set S with N elements. Initialize the set f(yj): y = f(xi) is the points xi with the same value. If j = 1: End Calculate the standard deviation of each group f(yj) (yj is the jth element in the set f(y)) $\sigma(f(y_j)) = Sqrt(1/N^* \Sigma (x_i - \mu)^2)$ Calculate the standard deviation of the set f(y) and f(x). $\sigma(f(y)) = Sqrt(1/N^* \Sigma(y_i - \mu_y)^2)$ … $\sigma(f(x)) = \sigma\, j \to n\, (j/N^* \sigma(f(y_j))$ Partition by RSS: $RSS(s_k) = \sigma(f(y)) - \sigma(f(x))$ Go back to step 2 with the next k. Evaluating the new model with the largest RSS(sk) will have the most sigma reduction. - End

F. Evaluation

To develop predictive models, we use data from 2017, 2018, and 2019.

Fig. 5. Evaluation methodology illustration.

And, to assess the accuracy of the prediction results, it is advised to research using the MSE and RMSE formulae to determine the overall mean error with a big quantity of data and the prediction results having a substantial departure from reality. To test the accuracy, the user compares the measured value to the real business value of the firm in 2019. The prediction model's efficacy may then be validated, is shown through the evaluation methodology (see Fig. 5) and flowchart (see Fig. 6).

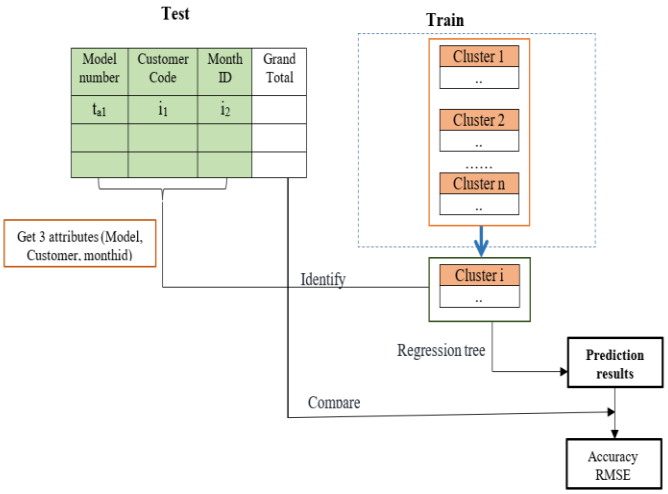

Fig. 6. Flowchart of the test protocol execution process.

V. Results

A. Results of Clustering on the Data Set

Before employing clustering methods, the number of groups k appropriate for clustering must be determined. Choosing k becomes simple for data sets with a modest number of components. Choosing k in a data collection with a high number of components is similarly tough. Combine using the Elbow approach, and then proceed to pick the best k possible using the Elbow chart, which shows the relationship between the SSE value and the number of k groups, with k ranging from 1 to 9 (see Fig. 7). The graph shows a considerable bend at the value k = 3, indicating that the value k = 3 is the optimal number of clusters.

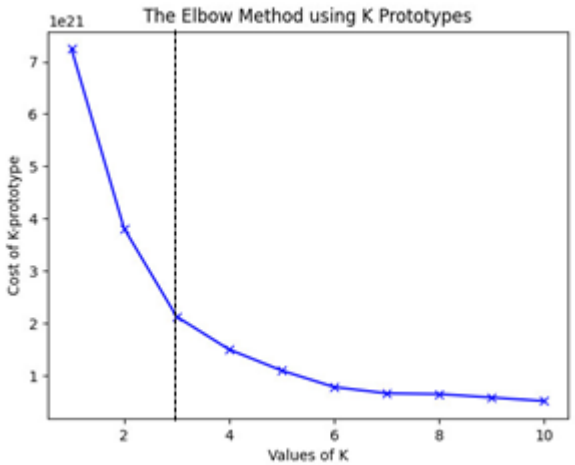

Fig. 7. Elbow method results on specific data set.

After determining the value of K=3, the study conducted clustering for the data set.

The number of items in each cluster can fluctuate as the centroids of the clusters change with the matching k centroids from Table II. With the data set comprising two types of characteristics, numeric attributes and discrete attributes, the result of the centre value of each cluster has two types

matching to the items in each cluster with k = 3. Fig. 8 depicts the information of the components in the appropriate cluster when k = 3.

TABLE II. INFORMATION ON CLUSTERS MATCHING TO THE CENTRE K

Cluster number k=3	Cluster	Number of elements
[[2.01743391e+05, '5000003626.0','RAC', 3.32668862e+09], [2.01777722e+05, '5000003903.0','REF', 1.83788563e+07], [2.01765167e+05, '5000003626.0','RAC', 4.00993996e+08]]	1	771422
	2	11874
	3	312

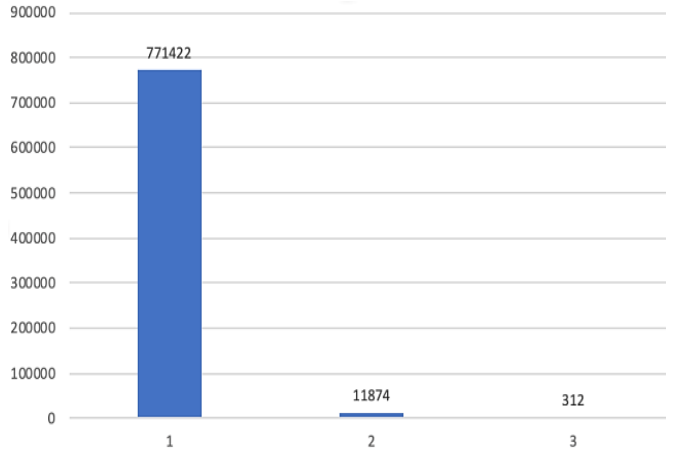

Fig. 8. The graph shows information of each corresponding cluster k=3.

B. Prediction Results on the Whole Data Set

For each cluster, the prediction results are presented on the regression tree.

Due to the big data set, when the tree partition is extremely large, the study reveals that the tree depth is five levels, and the findings mostly demonstrate the partitioning attribute at the root node for each cluster displayed via tree diagrams (Fig. 9, 10, 11).

Cluster 1:

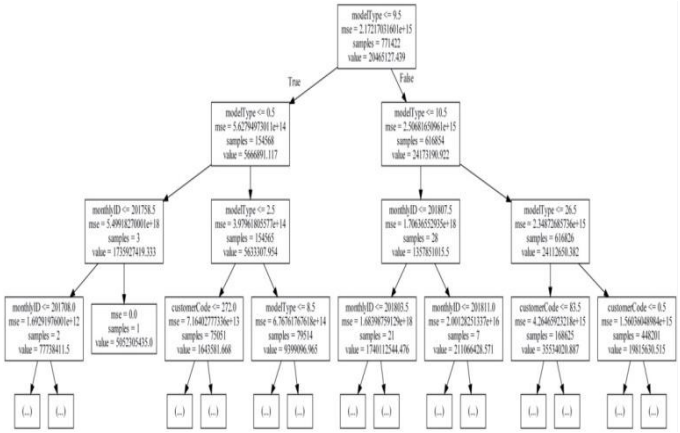

Fig. 9. The tree diagram displays information within the cluster.

Cluster 2:

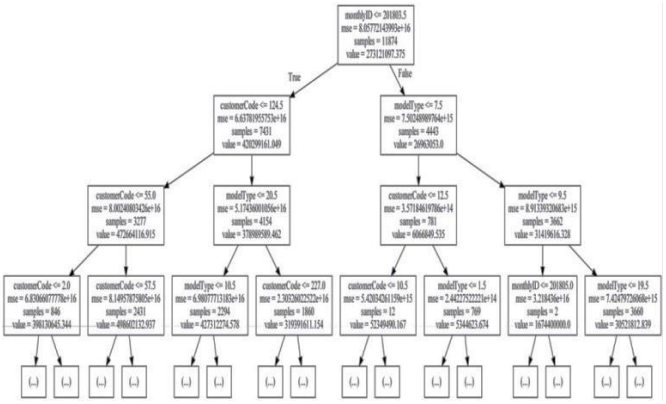

Fig. 10. The tree diagram displays information within the cluster.

Cluster 3:

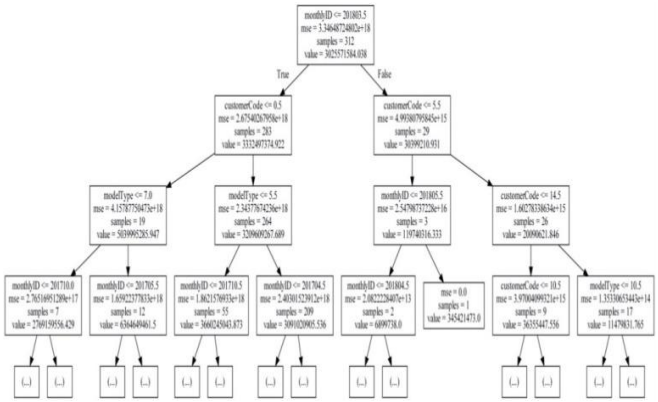

Fig. 11. The tree diagram displays information within the cluster.

The results of splitting the tree by cluster demonstrate that the data on each cluster is dispersed by distinct product groups.

Forecast results are exhibited quarterly throughout the year in the form of a column chart (Fig. 12, 13, 14, 15); the results reveal that the expected value is sometimes high and sometimes low owing to the real scenario caused by weather in each quarter and consumer wants changes, the value will change correspondingly.

Precious 1:

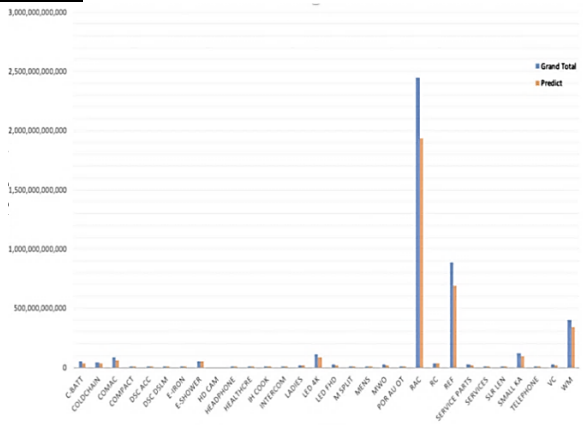

Fig. 12. Predicted and actual results on a specific Q1 data set.

Precious 2:

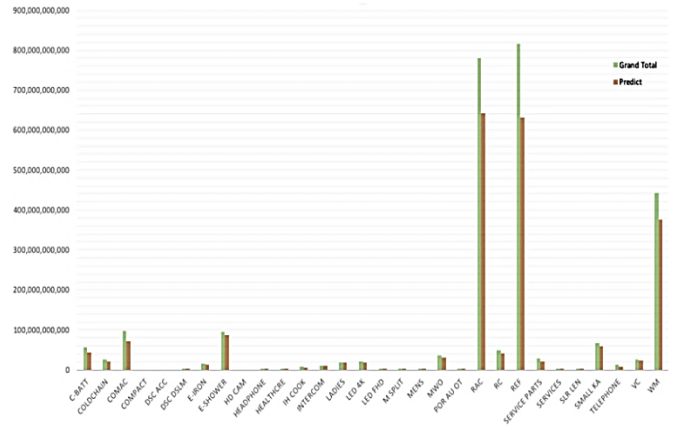

Fig. 13. Predicted and actual results on a specific Q2 data set.

Precious 3:

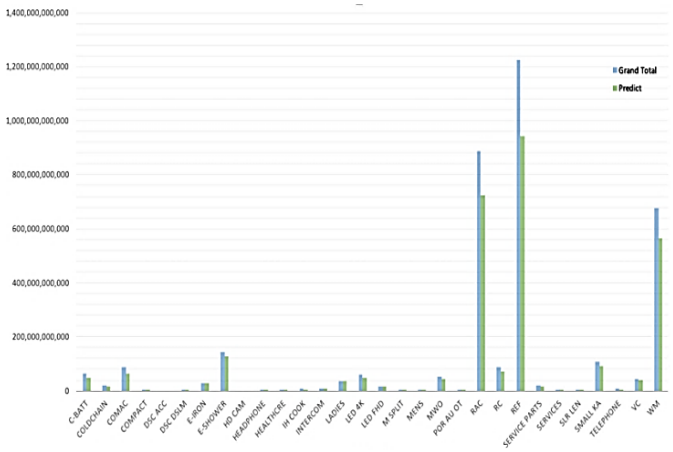

Fig. 14. Predicted and actual results on a specific Q3 data set.

Precious 4:

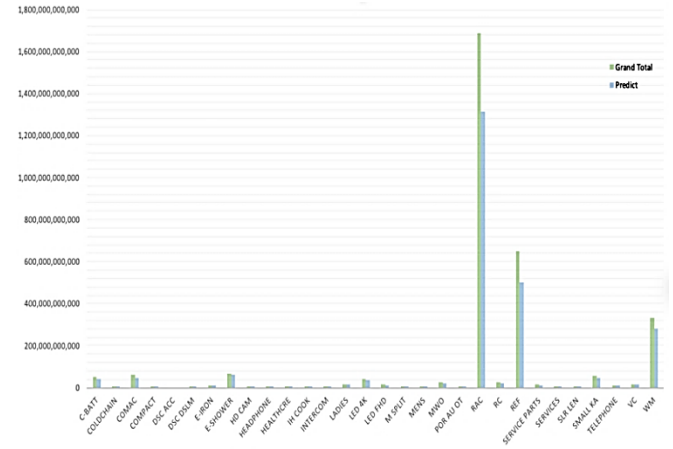

Fig. 15. Predicted and actual results on a specific Q4 data set.

Due to the enormous data collection, the research values are given on each quarter of the year to clearly highlight the difference of each separate product group.

The results of computing the standard deviation error of two cases when forecasting on a test data set based on real data.

TABLE III. DEVIATION RESULTS FOR EVALUATION INDICATORS

	RMSE	Execution time
When clustering is note performed	1.163.173,458	45 minute
After performing clustering	43.065,657	40 minute

According to the results of Table III, the deviation and implementation time are lower when clustering before predicting than when not clustering. As a result, the experimental model fits the requirements.

VI. CONLUSION

When the amount of information is really large, the sales forecasting approach is a very useful method for users. It titled "Application of data mining to forecast business trends for electronic products". Proposed clustering model paired with prediction algorithm to be used to the process in order to provide reliable prediction results. The data collection covers product information for three years. Divide the dataset into two parts: the training set contains data from 2017-2018, and the test set contains data from 2019. Using the training dataset with data characteristics of two types of mixed and numeric attributes, research will be conducted based on four attributes: product group (ModelType), customer code (Customer number), month code (Month id), and total quantity of goods sold based on the above three factors. With 3 clusters (k = 3), the number of elements in each cluster is 771422, 11874, and 312, respectively; Combined with the regression tree algorithm "Regression tree" to partition each cluster, using the evaluation protocol MSE, RMSE to evaluate the model's accuracy with over 80% results compared to the actual value, in order to build a system to predict product development trends/sales in the coming years. The experimental findings clearly indicate that using the sales prediction approach in a machine learning programming language yields results with an accuracy of more than 80%. The experimental time, whether rapid or slow, is determined by the original data collection. With the results of the experiment, it is feasible to apply a thorough test data set for each product group of particular categories on additional prediction models such as linear regression, Bagging, and so on in the future. Futures might be based on the outcomes of forecasts put into compact application.

REFERENCES

[1] Mehmet Yasin OZSAGLAM, “DATA MINING TECHNIQUES FOR SALES FORECASTINGS”, (September, 2015), PP. 6-9.

[2] Alfiah, Fifit et al. “Data Mining Systems to Determine Sales Trends and Quantity Forecast Using Association Rule and CRISP-DM Method.” (2018).

[3] Aditya Joshi et al, “Use of Data Mining Techniques to Improve the Effectiveness of Sales and Marketing”. IJCSMC, Vol.4 Issue.4, April-2015, pg. 81-87.

[4] Bhavin Parikh et al, “Applying Data Mining to Demand Forecasting and Product Allocations”. (2003).

[5] Ismail, Mustapha et al. “Data Mining in Electronic Commerce: Benefits and Challenges.” (2015).

[6] Ismail, Mustapha et al. “Data Mining in Electronic Commerce: Benefits and Challenges.” (2015).

[7] Do Thanh Nghi, Le Thanh Van Textbook of Knowledge Systems and Data Mining. Can Tho university, 2012.

[8] Do Thanh Nghi và Pham Nguyen Khang, 2012. Textbook of Principles of Machine Learning. Can Tho University, 137 pages.

[9] Do Thanh Nghi, Python Programming Language. Can Tho University, 2016.

[10] Breiman, Leo; Friedman, J. H.; Olshen, R. A.; Stone, C. J. (1984). Classification and regression trees. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software. ISBN 978-0-412-04841-8.

[11] Cosma Shalizi. Statistics 36-350: Data Mining, Fall 2006.

[12] Vu Huu Tiep, Text book Machine Learning co ban, June 15, 2019.

[13] Magidson, J & Vermunt, JK 2002, 'Latent class models for clustering: a comparison with K-means', Canadian Journal of Marketing Research, vol. 20, no. 1, pp. 36-43.

[14] Hendrickson, J. L.(2014). Methods for Clustering Mixed Data. (Doctoral dissertation). Retrieved from.

[15] L. Breiman, Bagging predictors, Mach. Learn, vol. 24, no. 2, pp. 123-140, 1996.

[16] Bühlmann, Peter. (2012). Bagging, Boosting and Ensemble Methods. Handbook of Computational Statistics. 10.1007/978-3-642-21551-3_33.

[17] Quinlan, J. R. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, 1993.

[18] NCSS Statistical Software, Chapter 446, [446-1:446-9].

[19] G. W. Milligan and M. C. Cooper. “An examination of procedures for determiningthe number of clusters in a data set”. Psychometrica, 50:159–179, 1985.

[20] T. Calinski and J. Harabasz, “A dendrite method for cluster analysis. Communications in Statistics”, 3:1–27, 1974.

[21] D T Pham, S S Dimov, and C D Nguyen, “Selection of K in K-means clustering”, 2004.

[22] Andrew Ng, “Clustering with the K-Means Algorithm, Machine Learning”, 2012.

[23] Robert L. Thorndike (December 1953). “Who Belong in the Family?”. Psychometrika 18 (4): 267–276.

[24] Karolis Urbonas, “Practical implementation of k-means clustering”.

[25] Peter J.Rousseeuw, “Silhouettes: a graphical aid to the interpretation and alidation of cluster analysis”, 1986.

[26] Prakash Nadkarni ,“Chapter 10 - Core Technologies: Data Mining and “Big Data””, 2016.

[27] Bednarz T. F., (2011): Sales Forecasting: Pinpoint Sales Management Skill DevelopmentTraining Series, Majorium Business Press, p.46.

[28] Kaufman L. and Rousseeuw P.J: Finding groups in Data. An introduction to cluster analysis. Wiley Interscience, 2005.

[29] Gatu, C., Yanev P.I., Kontoghiorghes, E.J., 2007. A graph approach to generate all possible regression submodels. Comput. Statist. Data Anal., 52, 799–815.

[30] Hofmann, M., Gatu, C., Kontoghiorghes, E.J., 2007. Efficient algorithms for computing the best subset regression models for large-scale problems. Comput. Statist. Data Anal., 52, 16– 29.

[31] Shih Y.S., Tsai H.W.,2004. Variable selection bias in regression trees with constant fits, Comput. Statist. Data Anal., 45, 595–607.

[32] A Maesya and T Hendiyanti, “Forecasting Student Graduation With Classification And Regression Tree (CART) Algorithm”, 2018.

[33] Barnston, A. G., 1992: Correspondence among the Correlation, RMSE, and Heidke Forecast Verification Measures; Refinement of the Heidke Score. Wea. Forecasting, 7, 699–709.

[34] Clustering Algorithms. By John A. Hartigan. New york: John Wiley and Sons, 1975.

[35] D T Pham, S S Dimov, and C D Nguyen, “Selection of K in K-means clustering”, 2004.

[36] Z. Huang, “Clustering large data sets with mixed numeric and categorical values,” in Proceedings of the 1st Pacific-Asia Conference on Knowledge Discovery and Data Mining, Singapore, 1997.

[37] Murty, M.. (2013). Cluster Analysis on Different Data Sets Using K-Modes and K-Prototype Algorithms. 249. 10.1007/978-3-319-03095-1_15.

[38] Özlem Akay & Güzin Yüksel (2018) Clustering the mixed panel dataset using Gower's distance and k-prototypes algorithms, Communications in Statistics - Simulation and Computation, 47:10, 3031-3041, DOI: 10.1080/03610918.2017.1367806.

[39] Dake, Delali & Gyimah, Esther & Buabeng-Andoh, Charles. (2023). University Students Behaviour Modelling Using the K-Prototype Clustering Algorithm. Mathematical Problems in Engineering. 2023. 10.1155/2023/5507814.

[40] Bitzenis, Aristidis & Koutsoupias, Nikos & Boutsiouki, Sofia. (2023). Business Research and Data Mining: a Bibliometric Analysis. 10.1109/ICECCME57830.2023.10252699.

[41] Alice, Dr & Andrabi, Syed & Jha, Shambhavi. (2023). Sales Forecasting Based on Ensemble Learning. 10.36227/techrxiv.24049452.v1.