P-ISSN : 2338-2899
E-ISSN: 2807-1271

Big Mart Sales Prediction Exploratory with the Concepts of Clustering

Dian Tri Wiyanti[1]*, Isnaini Rosyid*[2]

[1,2] *Department of Mathematics, Faculty of Mathematics and Natural Science, Universitas Negeri Semarang*

Correspondent Author: diantriwiyanti@mail.unnes.ac.id

Abstract — **Big marts record data related to product sales with their various dependent or independent factors as an important step to help forecast demand and inventory management in the future. Big mart tries to understand the properties of products and outlets that play an important role in increasing sales. While the increase in sales can not be separated from the relationship with customers. Customers can be grouped into different categories where marketing people can use targeted marketing and retain customers, this is commonly known as Customer Relationship Management (CRM). Managing a successful CRM implementation requires an integrated and balanced approach to technology, processes and people. Cluster analysis is widely used in many applications such as market research, pattern recognition, and data analysis, as it can help marketers find distinct clusters in their customer base.**

Keyword **—big mart, CRM, clustering.**

I. Introduction

Customer Relationship Management (CRM) is a company's approach to understanding and influencing customer behavior through good communication to improve customer acquisition, customer retention, customer loyalty, and customer profitability [1]. Customers could be clustered into different categories for which the marketing people can employ targeted marketing and retain the customers [2]. Therefore rules may be generated to increase business performance. At each customer touchpoint, the organization reinforces the value delivery while simplifying how customers interact and relate. To maximize the customer lifetime value, the sales representative strives to build a relationship and efficiently generate revenue from high potential prospects [3]. CRM is an active, participatory, and interactive relationship between business and customer with the objective is to achieve a comprehensive view of customers and be able to consistently anticipate and react to their needs with targeted and effective activities at every customer touchpoint [4]. To manage a successful CRM implementation requires an integrated and balanced approach to technology, process, and people. Big mart records data related to product sales with its various dependent or independent factors as an important step to assist in future demand prediction and inventory management. The dataset is built with data collected through customers, as well as data related to inventory management in the data warehouse, then refined to get accurate predictions [5]. Data mining is an iterative and interactive process to find a new pattern or model that is valid, usable, and understandable in a very large database, which contains a search for patterns or trends in a large database to assist future decision making [4]. There is a natural fit between data mining and CRM in that data mining techniques, when applied properly to the right data can be a powerful tool for formulating and implementing a good CRM strategy [6]. The CRM method that is often used to describe a set of market information is currently changing trends and CRM applications are supported by data from the data warehouse [7]. Data mining methods and applications can be used for decision-making in CRM in the areas of customer value and customer experience [4].

In this paper, sales data at Big Mart, a one-stop shopping center, has been used to build a predictive model and predict the sales of each product at a particular outlet. With the model obtained, BigMart will try to understand the properties of products and outlets that play an important role in increasing sales. It was reported that data may have missing values as some stores may not report all data due to technical glitches. Therefore, it is necessary to treat them accordingly.

The predictive power comes from a unique design by combining the process of extracting and discovering patterns in large data sets involving methods at the intersection of machine learning and statistics, while organizations get help to use their current reporting capabilities to discover and identify the hidden patterns in databases [2]. Clustering is a type of unsupervised learning where the goal is to partition a set of objects into groups called clusters, where these groups can be mutually exclusive or may overlap, depending on the approach used [8]. Cluster analysis is widely used in many applications such as market research, pattern recognition, and data analysis, as it can help marketers find distinct groups in their customer base. In addition, they can characterize their customer groups based on buying patterns. As a data mining function, cluster analysis serves as a tool to gain insight into the distribution of data to observe the characteristics of each cluster [9]. Data mining is a powerful technique to help companies find patterns and trends in their customer preferences, it is also a well-known tool for CRM [10].

The main data mining process uses data exploration technology to extract data, create predictive models using decision trees, and test and verify the stability and effectiveness of the model. The K-means method is to group customers into groups based on billing, loyalty, and payment behavior to create a decision tree-based model. Determining the number of k clusters in a data set with limited prior knowledge of the appropriate values is a general problem that is different from solving data clustering problems [11][12]. While X-Means is a data mining algorithm that is an extension of K-Means which is stated to be able to cover the shortcomings of K-Means [13]. We use K-Means and X-Means algorithms for predictive analysis of the Big Mart case. X-means clustering is used to overcome one of the main weaknesses of K-means clustering, it requires prior knowledge of the number of clusters (K) [14]. We use the two algorithms to show the difference in output, and how the results can provide input for decision-makers.

II. Methods

A. Algorithms and Research Data

This study uses the K-means and X-Means Algorithm which are the Data Mining algorithms used to cluster data. While the research data is data from data scientists at BigMart have earned sales input for 1,559 products in 10 stores in various cities. In addition, certain attributes of each product and shop have been defined. By these facts, the goal is to build a predictive model and predict the sales of each product in certain outlets. To evaluate the performance of the prediction and clustering under test, we use a collection of sales data from the online machine learning repository and data scientist community, Kaggle. The dataset contains a total of 8,522 samples with randomly selected training and testing data. Few of the training and testing data can be seen in Table 1 and Table 2.

A structured approach to data analytics is needed to discovering useful information from a collection of data in this research involves basic preprocessing up to results. The specific stages are described as follows:

1) Basic preprocessing
   At this stage, the data set is loaded and some basic preprocessing tasks are performed. Provides all labeled and unlabeled data points to which the model should be applied later, and normalizes the data and its normalized model so that the data can be changed later.
2) Feature engineering & modeling
   Performs selections using multi-purpose optimization and information preservation concepts, as well as performs actual grouping on the changed data.
3) Visualization
   Create visualization for the cluster model, which describes which data points belong to which cluster.
4) Process results

B. K-Means

The K-means algorithm is a simple iterative clustering algorithm. Using the distance as the metric and given the K classes in the data set, calculate the distance mean, giving the initial centroid, with each class described by the centroid. For a given data set X containing n multidimensional data points and the category K to be divided, the Euclidean distance is selected as the similarity index and the clustering targets minimizes the sum of the squares of the various types; that is, it minimizes [15]

$$\mathrm{d} = \sum_{k=1}^{k}\sum_{i=1}^{n} ||(x_i - u_k)||^2 \tag{1}$$

where k represents K cluster centers, $\mathcal{U}_k$ represents the kth center, and x_i represents the ith point in the data set. The solution to the centroid $\mathcal{U}_k$ is as follows:

$$\begin{aligned} \frac{\partial}{\partial u_k} &= \frac{\partial}{\partial u_k}\sum_{k=1}^{k}\sum_{i=1}^{n}(x_i - u_k)^2 \\ &= \sum_{k=1}^{k}\sum_{i=1}^{n}\frac{\partial}{\partial u_k}(x_i - u_k)^2 \\ &= \sum_{i=1}^{n} 2(x_i - u_k) \end{aligned} \tag{2}$$

Let Equation (2) be zero, then $\mathcal{U}_k = \frac{1}{k}\sum_{i=1}^{n} x_i$.

The central idea of algorithm implementation is to randomly extract K sample points from the sample set as the center of the initial cluster: Divide each sample point into the cluster represented by the nearest center point; then the center point of all sample points in each cluster is the center point of the cluster. Repeat the above steps until the center point of the cluster is unchanged or reaches the set number of iterations. The algorithm results change with the choice of the center point, resulting in an instability of the results. The determination of the central point depends on the choice of the K value, which is the focus of the algorithm; it directly affects the clustering results, such as the local optimality or global optimality [16].

C. X-Means

X-Means is a clustering algorithm that determines the correct number of centroids based on a heuristic. It begins with a minimum set of centroids and then iteratively exploits if using more centroids makes sense according to the data. If a cluster is split into two sub-clusters is determined by the Bayesian Information Criteria (BIC), balancing the trade-off between precision and model complexity In essence, the algorithm starts with K equal to the lower bound of the given range and continues to add centroids where they are needed until the upper bound is reached. During this process, the centroid set that achieves the best score is recorded, and this is the one that is finally output. The algorithm consists of the following two operations repeated until completion :

1) The improve-params operation: consists of running conventional K-means to convergence.

TABLE 1
A MINOR PART OF THE DATA TRAINING USED

Item_ Identifier	Item_ Weight	Item_Fat _Content	Item_Visibility	Item_Type	Item_MRP	Outlet_Identifier	Outlet_Establishment_ Year	Outlet_ Size	Outlet_Location_Type	Outlet_Type	Item_Outlet_ Sales
FDA15	9.3	Low Fat	0.016047301	Dairy	249.8092	OUT049	1999	Medium	Tier 1	Supermarket Type1	3735.138
DRC01	5.92	Regular	0.019278216	Soft Drinks	48.2692	OUT018	2009	Medium	Tier 3	Supermarket Type2	443.4228
FDN15	17.5	Low Fat	0.016760075	Meat	141.618	OUT049	1999	Medium	Tier 1	Supermarket Type1	2097.27
FDX07	19.2	Regular	0	Fruits and Vegetables	182.095	OUT010	1998		Tier 3	Grocery Store	732.38
NCD19	8.93	Low Fat	0	Household	53.8614	OUT013	1987	High	Tier 3	Supermarket Type1	994.7052
FDP36	10.395	Regular	0	Baking Goods	51.4008	OUT018	2009	Medium	Tier 3	Supermarket Type2	556.6088
FDO10	13.65	Regular	0.012741089	Snack Foods	57.6588	OUT013	1987	High	Tier 3	Supermarket Type1	343.5528
FDP10		Low Fat	0.127469857	Snack Foods	107.7622	OUT027	1985	Medium	Tier 3	Supermarket Type3	4022.7636

TABLE 2
A MINOR PART OF THE DATA TESTING USED

Item_ Identifier	Item_ Weight	Item_Fat _Content	Item_Visibility	Item_Type	Item_MRP	Outlet_Identifier	Outlet_Establishment_ Year	Outlet_ Size	Outlet_Location_Type	Outlet_Type	Item_Outlet_ Sales
FDW58	20.75	Low Fat	0.007564836	Snack Foods	107.8622	OUT049	1999	Medium	Tier 1	Supermarket Type1	FDW58
FDW14	8.3	reg	0.038427677	Dairy	87.3198	OUT017	2007		Tier 2	Supermarket Type1	FDW14
NCN55	14.6	Low Fat	0.099574908	Others	241.7538	OUT010	1998		Tier 3	Grocery Store	NCN55
FDQ58	7.315	Low Fat	0.015388393	Snack Foods	155.034	OUT017	2007		Tier 2	Supermarket Type1	FDQ58
FDY38		Regular	0.118599314	Dairy	234.23	OUT027	1985	Medium	Tier 3	Supermarket Type3	FDY38
FDH56	9.8	Regular	0.063817206	Fruits and Vegetables	117.1492	OUT046	1997	Small	Tier 1	Supermarket Type1	FDH56
FDL48	19.35	Regular	0.082601537	Baking Goods	50.1034	OUT018	2009	Medium	Tier 3	Supermarket Type2	FDL48
FDC48		Low Fat	0.015782495	Baking Goods	81.0592	OUT027	1985	Medium	Tier 3	Supermarket Type3	FDC48

2) The improve-structure operation: finds out if and where new centroids should appear. This is achieved by letting some centroids split in two. It begins by describing and dismissing two obvious strategies, after which we will combine their strengths and avoid weakness in the X-means strategy.
3) If $K > K_{max}$ stop and report the best-scoring model found during the search. Else, go to 1 [17].

III. Results and discussion

This stage consists of an evaluation of the pattern to identify useful patterns representing knowledge based on some appropriate and appropriate actions, and knowledge presentation, to present mined knowledge to decision-makers.

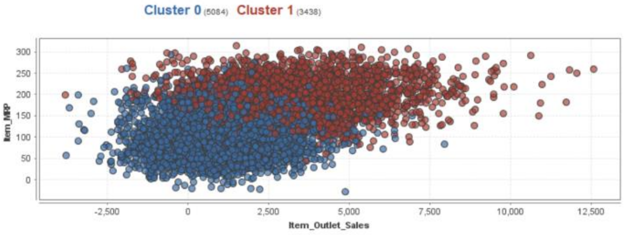

Figure 1. K-Means scatter plot.

Figure 1 shows the output of K-Means gives 2 clusters, that is clusters 0 and 1. In cluster 0 (blue dots), the attribute item_outlet_sales is on average 45.59% smaller, item_MRP is on average 33.83% smaller, and item_weight is on average 2.62% smaller. While in cluster 1 (red dots), the attribute item_outlet_sales is on average 67.42% larger, item_MRP is on average 50.02% larger, and item_weight is on average 3.87% larger. From the Figure 1 we could see that the two most important attributes as part of future prediction decisions are item_outlet_sales and item_MRP.

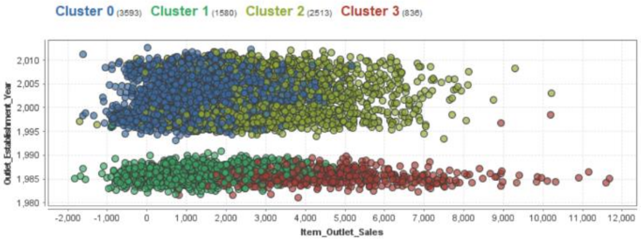

Figure 2. X-Means scatter plot.

Meanwhile, Figure 2 represents the result of X-Means which gives 4 clusters, that is clusters 0 (blue dots), 1 (dark green dots), 2 (light green dots), and 3 (red dots). In cluster 0, the attribute item_outlet_sales is on average 41.70% smaller, outlet_establishment_year is on average 38.71% larger, and item_MRP is on average 32.77% smaller.

While cluster 1 gives results with the attribute item_outlet_sales is on average 44.27% smaller, outlet_establishment_year is on average 91.44% smaller, and item_MRP is on average 18.44% smaller.

Then cluster 2 brings out an average of 49.31% smaller for the attribute item_outlet_sales, outlet_establishment_year is on average 34.18% larger, and item_MRP is on average 47.11% larger.

The last cluster is cluster 3 shows the attribute item_outlet_sales is on average 114.70% larger, outlet_establishment_year is on average 96.30% smaller, and item_MRP is on average 34.07% larger.

The Figure 2 also indicates that the two most important attributes as part of future prediction decisions are item_outlet_sales and outlet_establishment_year.

For the final result, the centroid table of K-Means and X-Means can be seen in Table 3 dan 4.

The item_outlet_sales attribute is sales of the product in the particular store, and item_MRP is the maximum retail price (list price) of the product. While item_weight is the weight of the product, and the outlet_establishment_year attribute is the year in which the store was established. The summary of statistics obtained is shown in Table 5.

IV. Conclusion

In this paper, the main objective to be conveyed is the use of clustering in a CRM system which is a fascinating and effective technique for customer clustering so that it can produce irresistible information. Thus, BigMart will try to understand the properties of products and outlets that play an important role in increasing sales from sales data for 1559 products in 10 stores in various cities. In the implementation, the results of clustering show that not all attributes affect the high sales results. Several important attributes need to be considered by decision-makers, and how certain outlet locations record the highest sales, other shopping locations need to follow the same pattern to increase sales.

The overall goal of the data mining process is to extract information from a large data set and convert it into an understandable form for further use. Clustering is important in data analysis and data mining applications by grouping a set of objects so that objects in the same group are more similar to each other than those in other groups (clusters).

Another interesting future work is on the use of data mining classification techniques in CRM systems in order to not only analyze customer behavior but also to predict it. In addition, it is quite interesting to integrate clustering and classification algorithms in business intelligence systems to make it easier for marketing and sales teams to use them.

TABLE 3
K-MEANS CENTROID TABLE

Cluster	Item_Fat _Content	Item_MRP	Item_Outlet _Sales	Item_Type	Item_ Weight	Outlet_ Identifier	Outlet_ Location_ Type	Outlet _Size	Outlet _Type
Cluster 0	1	103.889	1202.132	13	12.640	3	2	1	1
Cluster 1	1	195.880	3629.646	6	13.179	5	1	1	1

TABLE 4
X-MEANS CENTROID TABLE

Cluster	Item_Fat _Content	Item_MRP	Item_Outlet _Sales	Item_Type	Item_ Weight	Outlet_ Establishment _Year	Outlet_ Identifier	Outlet_ Location _Type	Outlet _Size	Outlet _Type
Cluster 0	1	105.049	1285.568	6	12.194	2002.799	7	1	3	1
Cluster 1	1	120.774	1230.366	13	12.965	1986.099	1	2	0	1
Cluster 2	1	192.684	3240.694	13	13.718	2002.218	9	0	1	1
Cluster 3	1	178.381	4645.302	6	12.917	1985.475	5	2	1	3

TABLE 5
DATA TESTING

Attribute Name	Value			
	item_outlet_sales	item_MRP	item_weight	outlet_establishment_year
Minimum	33.290	31.290	4.555	1985
Maximum	13086.965	266.888	21.350	2009
Average	2181.455	141.000	12.857	1997.832
Standard Deviation	1706.531	62.275	4.644	8.372

DAFTAR ACUAN

[1] R. Swift, *Accelerating Customer Relationship Using CRM and Relationship Technologies*. New York: Prentice Hall Inc., 2001.

[2] I. Enesi, L. Liço, A. Biberaj, and D. Shahu, “Analysing Clustering Algorithms Performance in CRM Systems,” in *Proceedings of the 23rd International Conference on Enterprise Information Systems (ICEIS 2021)*, vol. 1, no. 1, pp. 803–809, 2021.

[3] C. Fisher, “New Technologies for Mobile Salesforce Management and CRM,” *Am. J. Ind. Bus. Manag.*, vol. 7, no. 4, pp. 548–558, 2017.

[4] A. Dwiastuti, A. Larasati, and E. Prahastuti, “The implementation of Customer Relationship Management (CRM) on textile supply chain using k-means clustering in data mining,” *MATEC Web Conf.*, vol. 204, 2018.

[5] N. Malik and K. Singh, “Sales Prediction Model for Big Mart,” *Parichay Maharaja Surajmal Inst. J. Appl. Res.*, vol. 3, no. 1, pp. 22–32, 2020.

[6] R. S. Winer, “A framework for customer relationship management,” *Calif. Manage. Rev.*, vol. 43, no. 4, pp. 89–105, 2001.

[7] A. Khan, N. Ehsan, E. Mirza, and S. Z. Sarwar, “Integration between Customer Relationship Management (CRM) and Data Warehousing,” *Procedia Technol.*, vol. 1, pp. 239–249, 2012.

[8] A. Cornuéjols, C. Wemmert, P. Gançarski, and Y. Bennani, “Collaborative clustering: Why, when, what and how,” *Inf. Fusion*, vol. 39, pp. 81–95, 2018.

[9] Tutorials Point, *“Data Mining - Cluster Analysis”* https://www.tutorialspoint.com/data_mining/dm_cluster_analysis.htm.

[10] H. I. Arumawadu, R. M. K. T. Rathnayaka, and S. K. Illangarathne, “Mining Profitability of Telecommunication Customers Using K-Means Clustering,” *J. Data Anal. Inf. Process.*, vol. 03, no. 03, pp. 63–71, 2015.

[11] R. M. K. T. Rathnayaka, “Cross-Cultural

Dimensions of Business Communication: Evidence from Sri Lanka," *Int. Rev. Manag. Bus. Res.*, vol. 3, no. 3, pp. 1579–1588, 2014.

[12] R. M. K. T. Rathnayaka, D. M. K. . Seneviratna, and W. Jianguo, "Grey system based novel approach for stock market forecasting," *Grey Syst. Theory Appl.*, vol. 5, no. 2, pp. 178–193, 2015.

[13] A. Radwan *et al.*, "X-means clustering for wireless sensor networks," *J. Robot. Netw. Artif. Life*, vol. 7, no. 2, pp. 111–115, 2020.

[14] M. Mughnyanti, S. Efendi, and M. Zarlis, "Analysis of determining centroid clustering x-means algorithm with davies-bouldin index evaluation," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 725, no. 1, 2020.

[15] Q. Wang, C. Wang, Z. Feng, and J. Ye, "Review of K-means clustering algorithm," *Electron. Des. Eng*, vol. 20, pp. 21–24, 2012.

[16] R. R. Rathod and R. D. Garg, "Design of electricity tariff plans using gap statistic for K-means clustering based on consumers monthly electricity consumption data," *Int. J. Energy Sect. Manag.*, vol. 11, no. 2, pp. 295–310, 2017.

[17] D. Pelleg and A. W. Moore, "X-means: Extending k-means with efficient estimation of the number of clusters," in *In Icml*, pp. 727–734, 2000.