Effective Analysis of Sales Data Set Using Advanced Classifier Techniques

G.S. Ramesh, Asst. Prof., Department of CSE, VNR VJIET. E-mail: ramesh_gadwal@yahoo.com

Dr.T.V. Rajini Kanth, Professor, Department of CSE, Sreenidhi Institute of Science and Technology. E-mail: rajinitv@gmail.com

Dr.D. Vasumathi, Professor, Department of CSE, JNTUH-CEH. E-mail: vasukuar_devaara@yahoo.co.in

G.V.S. Akash Bharadwaj, B. Tech. Student, SNIST. E-mail: gvsakashb@gmail.com

Abstract--- Business data is growing day by day very rapidly and it is very important to visualize the hidden patterns of the large business data sets effectively to enhance the business and also to compete with the upcoming business competitors apart from full filling the dynamic requirements of customers. The Machine learning is one of the upcoming technologies or methods to address the business challenges and helps the business growth. In this paper, various classifiers were applied on the pre-processed and K- means clustered data set. Comparison was made and summarized the outcomes with the purpose to know the suitability of the classifier procedure for effectiveness of analysis of the data set. It was found that NNGE classifier was the best among all and followed by M5P tree classifier. Apart from that, Explorative data analysis was also made on the pre-processed data set in order to exploit the hidden patterns for effective conclusions.

Keywords--- Business Data, NNGE Classifier, M5P Classifier, K-means Cluster, Rider Rules.

I. Introduction

Business Intelligence is a collection of techniques or methods that help organizations to analyze growing structured and un-structured complex data effectively and make them to take effective decisions. Particularly the Product sales data will be analyzed efficiently and identifies suitable patterns for predictive models in order to understand predict customer buying patterns. The Machine Learning algorithms are playing major in prediction and classification of huge voluminous sales data apart from constructing suitable models.

II. Related Work

S. Hanumanth Sastry et. al. [1] stated that clustering techniques will be useful in detect deviation in sales of products apart from comparison of certain period sales. They have used clustering techniques namely K-means and Expected Maximization on steel product sales and found interesting rules to improve steel sales. The attributes they have considered are customers, Products and sold quantities and also on factors of demand like customer profile, discount, price and sales tax etc. The two clustering techniques K-means, EM are found to be better than other clustering techniques namely DBSCAN, OPTICS and COBWEB. Kiran Singh et. al. [2] stated that huge data sets were created using Remote Satellites, Traffic Cameras, Social Network sites like Face book, Twitter etc., Health care data base, Government data and Business data and are throwing challenges in terms of analysis and prediction of useful patterns. It is very difficult to visualize the huge voluminous data so they focused on a system in order to enhance business revenue, tracking of progress and to make effective decision making. Qi Zhang et.al. [3] stated that Industry 4.0 or IOT uses Information and Communication technology in order to process huge data sets pertaining to products, machines and human resources together. Based on sales data and customer feedback product need to be innovated or reformed and brings in to production stage. They mainly focused on car sales data and analyzed using Data Mining Technology using Java language developed Web crawler application and to suggest industry in terms of optimization of resources apart from wastage reduction. Fatimetou Zahra Mohamed Mahmoud [4] stated that the usage of analytical systems is increasing by the existence of big data in order to analyze effectively. Predictive Analytics are very helpful in predicting the unknown patterns based on the hidden patterns in the large voluminous data to make effective preventive decisions. Discussion was made on the type of Predictive Analytics and its purpose to solve the problem of that industry which varies across industries. It was found that these Predictive Analytics are useful in optimum usage of Resource management apart from taking better strategic decisions in preventing risk and reduce cost etc. The research in this area show challenges in cleaning the data and in the development of suitable predictive models. Merve Turkmen Barutcu [5] stated that the marketing domain is generating huge massive data sets across 24 x 7 basis and by which we will know about customer behavior and their purchasing patterns. The analysis on customer opinions from social media, and money spent on product promotions

DOI: 10.5373/JARDCS/V11I11/20193195
ISSN 1943-023X

Received: 16 Sep 2019/Accepted: 16 Oct 2019

will provide us decisions what innovations required to enhance product and improve sales and profit. Big data will make us to understand the transactions with in industries and make us to predict the future marketing. It is required to explore challenges and opportunities growing in marketing industry using these technologies and enhance the competitiveness to with stand in the marketing. Juan Rincon-Patino et. Al. [6] stated that the fruit named Persea Americana popularly known as Avacado sales were increasing day by day in international market. Although there are verities of Avacados vailable out of which Hass variety was sold more. The information extracted from the Avacado sales data will be analyzed for effective business decision making. Researchers started analyzing the sales data of Hass variety of Avacado in terms of Weather and legacy sales data in several cities of USA in order to find units of sales and other information. They have used 4 algorithms namely Multivariate Regression prediction Models, Support vector Machines for regression, Multilayer Perceptron and Linear Regression out of which the first two were shown good results than the other last two methods. The Multivariate Regression Prediction Model made manufacturer and vendors avocado to device trades through the approximation of the returns in dollars and the quantity of avocados that could be traded in USA. Alessandro Massaro et. al. [7] stated that they have studied the 3 years large chain of retail stores data set by the application of Rapid minor workflows. They have constructed a deep learning model to predict the sales forecasting. This model basically depends on ANN and was applied on pre-processed historical data.

This model used multi layered neural network along with an optimized operator in order to discover finest parameter to implement the procedure. The results were compared with different Data Mining Algorithms like SVM, KNN, Deep learning, Decision Trees and Gradient Boosted Trees Deep learning based on the parameters namely degree of correlation amid actual and forecast values, Mean Absolute error, Relative average error. The ANN showed to be the best in terms of performance. Aditya Joshi et. al. [8] stated that extraction of patterns and classification from customer data are required for effective decision making. It is required to identify new trends time to time and sales patterns from inventory indicates market trends and use this to forecast sales patterns and helpful for strategic planning. Initially divide the stock data in to 3 clusters using k-meaans clustering technique on the basis of parameters like Product categories, solid quantities, Dead stock, slow moving and fast moving. After that use maximum recurrent configuration procedure to find recurrent configurations of attributes to find sales trend. Anurag Bejju [9] stated that in the present world E-commerce is the strongest catalyst for the growth of economic development in terms of decreasing operational cost, reducing financial barriers etc. Many companies are restructuring business to attain maximum customer satisfaction. E-commerce is not only for product trading apart from that it provides opportunity to compete with other business giants. Data Mining (DM) techniques are helping the business community to take effective decision making to enhance their profit. These DM techniques will reveal customer buying patterns. The author's objective is to improve traditional pricing strategies with the help of derived useful information from DM techniques. The proposed strategy is developed by the optimization of decision trees. Abhilash C B et.al. [10] Stated that Machine learning methods are mostly used in clustering of records. The most popular method is K-means clustering. It was found from experimental work is that clustering is resulted with less correctness by more number of repetitions. An enhanced version of k-means clustering was replicated with the help of least spanning tree.

In this approach an undirected plot was produced for every intake point and Least distance is identified and this outcomes in less number of iterations and better accuracy. The comparison results were analyzed and found that the enhanced k-means clustering technique has superior performance. The quantity of cluster points raises the correctness of the procedure using enhanced k-means algorithm. This also showed that at a specific value of k (cluster groups) the correctness of proposed procedure was optimum. Lipika Dey et.al. [11] Stated that Predictive analytics based on time series analysis on the given data set is always throws challenges on researchers. Researchers are trying to develop efficient predictive data analytic techniques on Huge textual information consisting of both structured and unstructured which is generating across web. They have tried to address this problem by developing deep Learning Frame work for predictive analytics. The frame work was tested using LSTM for forecasting of direction of information movement using news objects. Investigational outcomes showed that the proposed model overtakes existing methods. Marko Bohanec et. al. [12] stated that decision makers sometimes forced to take verdicts based on subjective prototypes, reflecting their knowledge. Investigation has exposed that data driven decision making approach performed better. This has made companies to present intellectual data driven decision prototypes for commercial situation. A novel wide-ranging explanation method recommended and it supported Black-Box prediction model.

They have proposed the explanation methodology in B2B sales forecasting. Users can validate explanations and test their hypothesis. Results demonstrated the successfulness of this method and it also made to make general suggestions in sales strategy. The flexibility of this method made suitable for many different Applications. They

DOI: 10.5373/JARDCS/V11I11/20193195
ISSN 1943-023X

Received: 16 Sep 2019/Accepted: 16 Oct 2019

have recorded a real-world circumstance to resolve decision support challenge which can be extended for the use of intelligent systems. There is a separation between Machine learning prototype selection and Model explanation. Clarifications unrelated to a particular forecast prototype positively affect acceptance of novel and complex prototypes.

III. Experimental Results

The data set consists of 13 attributes and in that Fig.1 is showing Product Type vs Product Cost. The highest product cost is for the product Eye wear followed by watches, Lanterns, Navigation, Cooking Gear, Woods, Irons, Tents, …, Sunscreens and lastly putters.

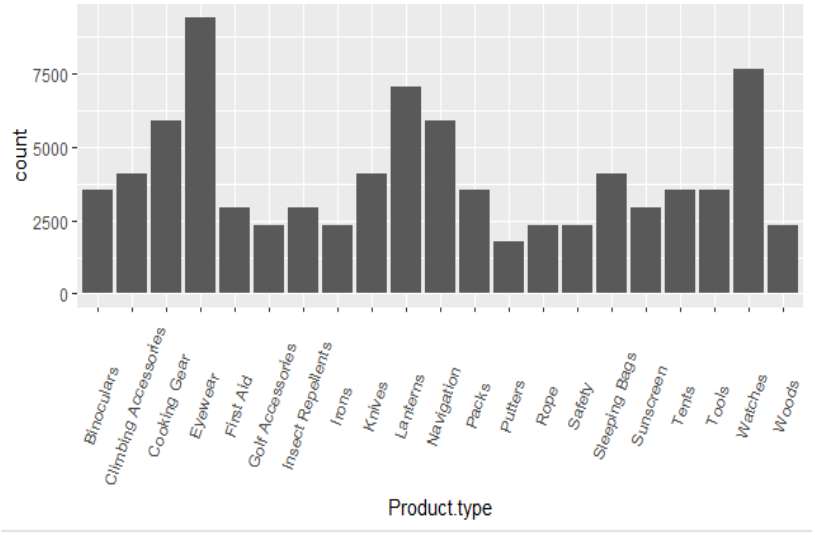

Fig. 1: Products Type vs. Products Cost

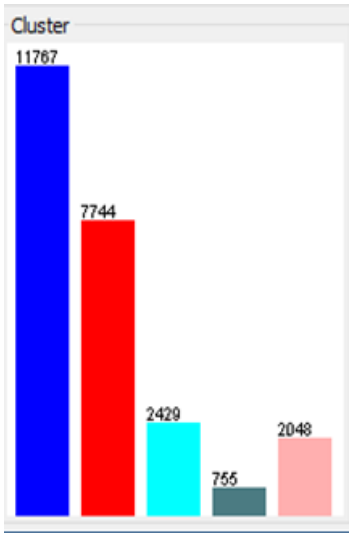

Fig. 2: Cluster Histogram

Fig. 3: shows the Retailer country vs Revenue in which it observed that the revenue of United states has highest Revenue touching near to Rs. 60,00,000/- when compared to other remaining 20 countries. Denmark has lowest revenue in between Rs.35,00,000 to Rs.37,50,000/- .

The second highest revenue around Rs.55,00,000/- touched by the country Japan followed by China and United kingdom.

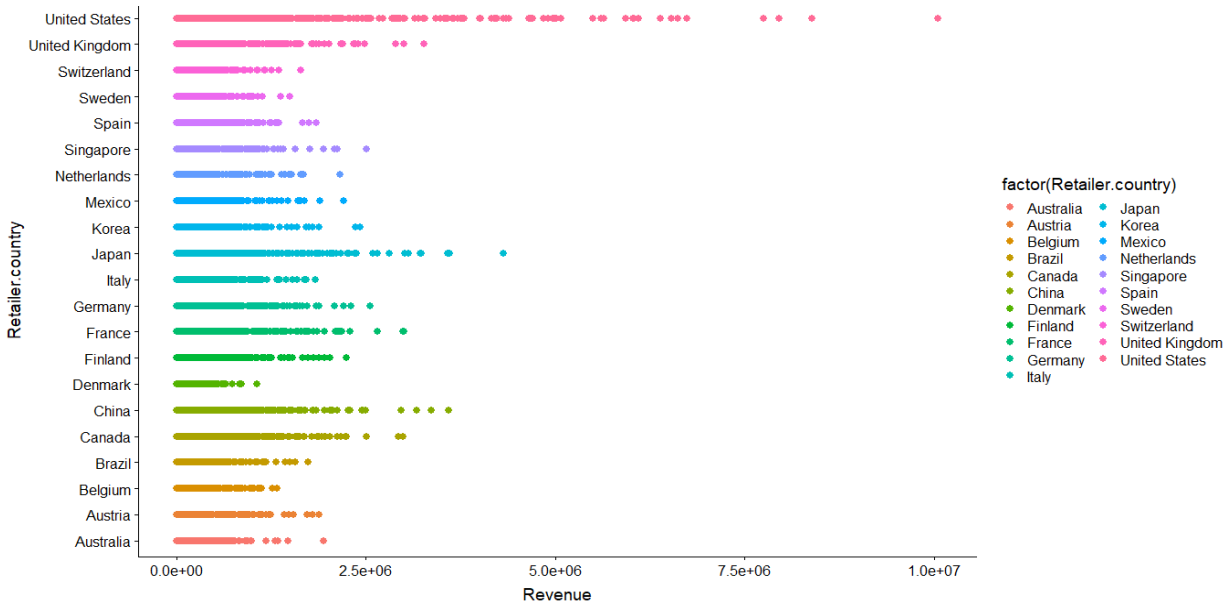

Fig. 3: Retailer Country vs. Revenue

DOI: 10.5373/JARDCS/V11I11/20193195
ISSN 1943-023X
Received: 16 Sep 2019/Accepted: 16 Oct 2019

Fig. 4: Shows the Product line shows there are 5 types namely Camping Equipment, Personal Accessories, Outdoor Protection, and Mountaineering equipment followed by Golf equipment. Across all these five major share in Product line goes to camping Equipment followed by Personal accessories.

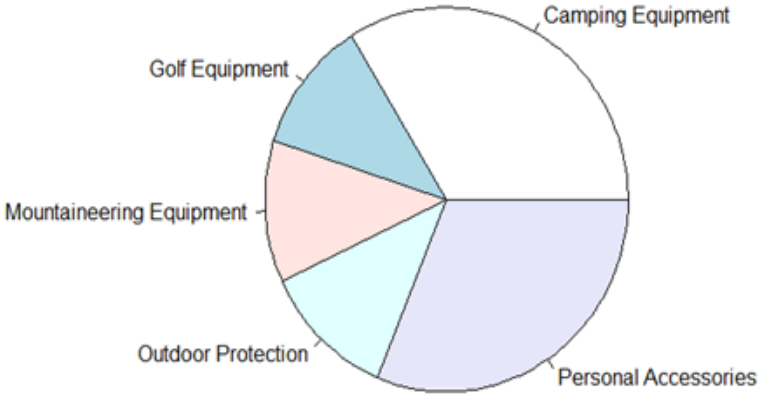

Fig. 4: Product Line Pie Chart

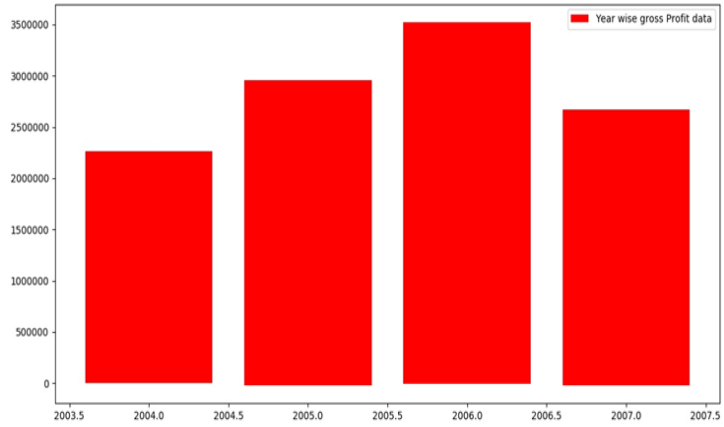

Fig. 5: Year Wise Gross Profit

Fig. 5: shows Year wise gross profit year 2006 has highest gross profit followed by 2005, 2007, and 2004. The linear regression models was constructed to predict Revenue based on the attributes Product.cost, Quantity, Unit.cost, Unit.Price, Gross.profit, Unit.sale.Price after removal of Null values

Revenue ~ Product.cost + Quantity + Unit.cost + Unit.price + Gross.profit + Unit.sale.price (1)

The Data set is subjected CfsSubsetEval method to select significant attributes and found the 5 attributes out of 14 attributes and they are namely Product, Unit cost, Unit price, Gross profit and Unit sale price. A new data set was formed using these 5 attributes and removing other attributes. k-means Clustering Technique was applied over that reduced data set for given k =5 clusters with Euclidean distance for centroid calculation. Table-1 shows the Clustered Data set of 5 clusters formed in which Cluster- 0 has 48%, Cluster-1 has 31%, Cluster-2 has 10%, Cluster-3 has 3% and Cluster-4 has 8% of Instances. Most of the instances were categorized under Cluster-0 followed by Cluster-1 and with least by Cluster-3.

Cluster -0 has the product name Star Peg with low values across all the attributes like Unit Cost, Unit Price, Gross Profit and Unit Sale Price when compared with all other clusters. Cluster-1 has the product name Polar Sun who's Unit Cost, Unit Price, Gross Profit; Unit Sale Price attribute values are higher than Cluster-0and Cluster-3 but lower than Clutser-2 and Cluster-4. Cluster-2 has the product Hibernator Extreme whose attributes namely Unit Cost, Unit Price, Gross Profit; Unit Sale Price has values higher than Cluster-0, Cluster-1and Cluster-3 but lower than Cluster-4. Cluster-4 has the product Star Dome whose attributes namely Unit Cost, Unit Price, Gross Profit; Unit Sale Price has values higher than Cluster-0, Cluster-1, Cluster-2 and Cluster-3. It has the highest Values when compared to all other clusters. Fig. 5 shows the Histogram of Clusters.

DOI: 10.5373/JARDCS/V11I11/20193195
ISSN 1943-023X

Received: 16 Sep 2019/Accepted: 16 Oct 2019

Table 1: Clustered Data Set

ATTRIBUTE	CLUSTER #					
	Full data (24743)	0(11767) 48%	1(7744) 31%	2(2429) 10%	3(755) 3%	4(2048) 8%
Product	Polar Sun	Star Peg	Polar Sun	Hibernator Extreme	Hawk Eye	Star Dome
Unit cost	84.8864	12.7241	62.2956	198.001	43.4658	466.0361
Unit price	155.9917	24.2079	113.4883	358.9661	79.6316	861.3006
Gross profit	77793.113	34113.2439	55443.1286	135357.3197	547416.8217	171870.2322
Unit sale price	147.2299	22.9346	106.7985	340.7345	78.1175	810.2381

Table 2: Performance Comparisons of Various Classifiers

Classifier Techniques	NB Tree	LAD Tree	Naive Baye's	Random Forest	Random Tree	J48	NNGE	M5 Pruned Tree	RIpple DOwn Rule Learner (Ridor)rules
% of Correctly Classified Instances	99.321	98.9492	96.9688	99.6848	99.4827	99.5676	99.7413	99.709	99.6565
Kappa Statistics	0.9897	0.984	0.9544	0.9952	0.9921	0.9934	0.9961	0.9956	0.9948
Precision	0.993	0.989	0.973	0.997	0.995	0.996	0.997	0.997	0.997
Recall	0.993	0.989	0.97	0.997	0.995	0.996	0.997	0.997	0.997
F-Measure	0.993	0.989	0.971	0.997	0.995	0.996	0.997	0.997	0.997
Mean Absolute Error	0.0036	0.0089	0.0126	0.0024	0.0021	0.0021	0.001	0.0036	0.0014
Relative Absolute Error	1.3479	3.3905	4.7952	0.8949	0.7834	0.7957	0.3928	1.3533	0.5217
Root Mean Square Error	0.0485	0.0609	0.0954	0.0325	0.0454	0.0402	0.0322	0.0343	0.0371
Time taken to create the model in Sec	4.44	49.58	0.06	6.69	0.03	0.27	2.58	31.7	20.81

The Table-2 and Fig. 6 are Showing performances of the various classifiers namely Naïve Baye's (NB) Tree, LAD Tree, Naïve Baye's, Random Forest, Random Tree, J48, NNGE, M5 Pruned Tree and Ripple Down Rule Learner (Ridor) rules were applied on the reduced clustered data set and compared. It is found that by and large the above classifiers proved to be good for the large sales Data set. The Best Classifier is found to be NNGE (Nearest-Neighbor-like algorithm using non-nested extensive examples (which are hyper rectangles that can be treated as if not then principles)) followed by M5 (Model Tree) Pruned Tree which then followed by Random Forest, Ridor, J48 and Random Tree in terms of Correctly Classified Instances and kappa statistic. Precision, Recall, F-measures are same for the classifiers NNGE, M5 Pruned Tree, Ridor and Random Forest next lower to these classifiers that followed by J48 and which then followed by Random Tree. Mean Absolute error was lowest for NNGE and increased slowly from Ridor, Random Tree, J48 and reaches highest for Naïve Baye's. Relative Absolute Error and Root Mean Square Error are lowest for NNGE. Time taken to execute was lowest for Random Tree, and increase s for Naïve Baye's, J48, NNGE etc. Considering all the factors it can conclude that NNGE is the Best Classifier followed by M5P, Ridor, Random Forest and J48 for large sales records set. The chart was displayed in below Fig. 6.

DOI: 10.5373/JARDCS/V11I11/20193195
ISSN 1943-023X

Received: 16 Sep 2019/Accepted: 16 Oct 2019

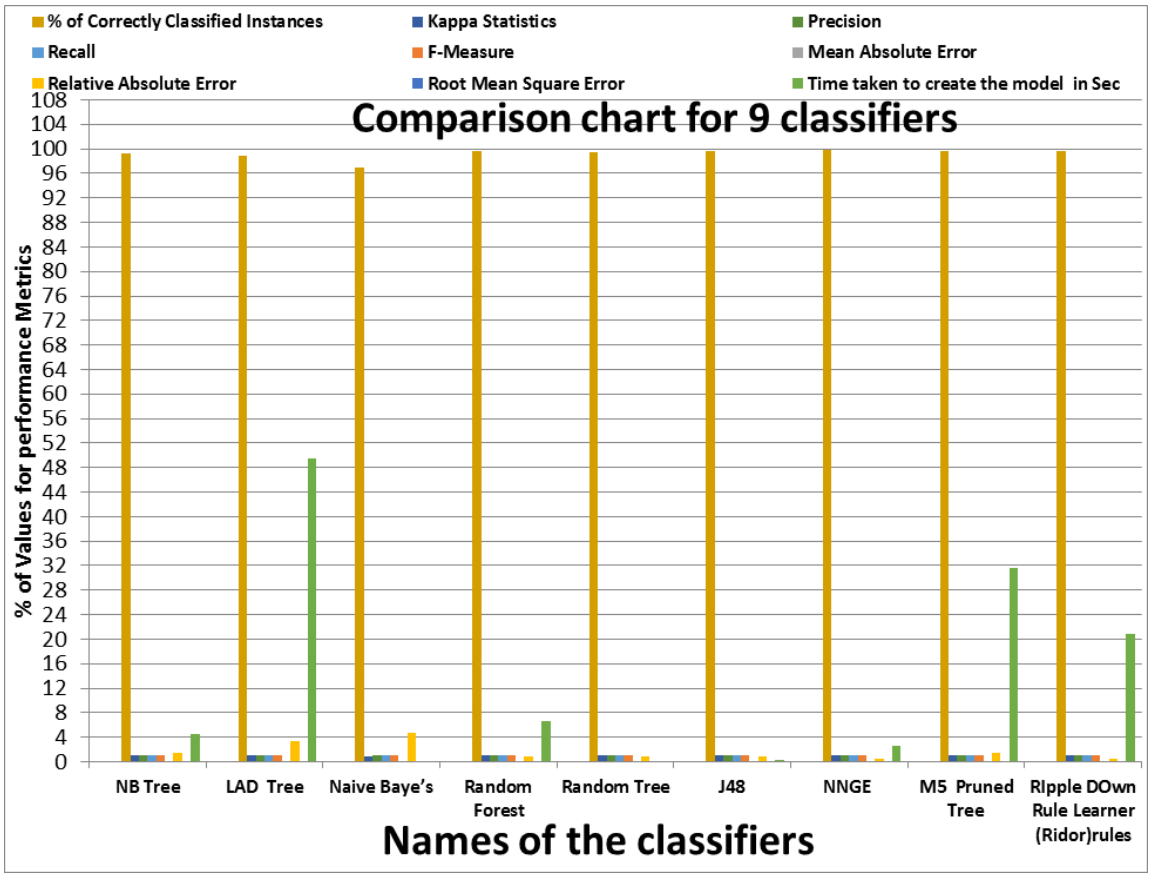

Fig. 6: Comparison Chart for Nine Classifiers

IV. Conclusion

The large sales Data set was classified using various Classifiers after preprocessing and attribute evaluation methods. The data set was reduced after CfsSubsetEval method for selection of Significant attributes by which 5 attributes were selected out of 14 attributes apart from removal of records where missing values are more. It was observed from the graph that highest product cost is for the product Eye wear followed by watches. It was found from histogram graphs that Revenue of USA is highest among all the countries followed by Japan, China, UK. The revenue of Denmark is the Lowest. There are 5 types of Product line in which the highest share was occupied by camping Equipment followed by Personal Accessories. The gross profit year 2006 was highest. The highest numbers of instances were clustered in Cluster-0 when compared to other clusters. Cluster -4 has highest values with respect to the attributes namely Unit Cost, Unit Price, Gross Profit; Unit Sale Price. It is observed that NNGE is the best classifier among all other classifiers followed by M5P, Ridor, Random tree and J48.

Acknowledgement

McKinley Stacker IV, SAMPLE DATA: Retail Sales & Marketing – Profit &Cost https://www.ibm.com/communities/analytics/watson-analytics-blog/retail-sales-marketing - profit-cost/

References

[1] S. Hanumanth Sastry and Prof. M.S. Prasada Bab, Analysis & Prediction Of Sales Data In SAP-ERP System Using Clustering Algorithms, *International Journal of Computational Science and Information Technology (IJCSITY)*, Vol.1, No.4, November 2013.

[2] Kiran Singh**,** Rakhi Wajgi, Data analysis and visualization of sales data, IEEE, 2016 *World Conference on Futuristic Trends in Research and Innovation for Social Welfare (Startup Conclave)* DOI: 10.1109/STARTUP.2016.7583967.

DOI: 10.5373/JARDCS/V11I11/20193195
ISSN 1943-023X
Received: 16 Sep 2019/Accepted: 16 Oct 2019

[3] Qi Zhang, Hongfei Zhan, Junhe Yu, Car Sales Analysis Based On the Application of Big Data, International Congress of Information and Communication Technology (ICICT 2017), *Procedia Computer Science* 107 (2017) 436 – 441, www.sciencedirect.com

[4] Fatimetou Zahra Mohamed Mahmoud. The Application of Predictive Analytics: Benefits, Challenges and How It Can Be Improved, *International Journal of Scientific and Research Publications,* Pg. No: 549 – 566, Volume 7, Issue 5, May 2017, ISSN 2250-3153.

[5] Merve Turkmen Barutcu, Big Data Analytics for Marketing Revolution, *Journal of Media Critiques [JMC]* doi: 10.17349/jmc117314 P-ISSN: 2056-9785 E-ISSN: 2056 9793 http://www.mediacritiques.net jmc@mediacritiques.net.

[6] Juan Rincon-Patino, Emmanuel Lasso and Juan Carlos Corrales, Estimating Avocado Sales Using Machine Learning Algorithms and Weather Data, Pg. No: 1-12, MDPI, Sustainability 2018, 10, 3498; doi: 10.3390/su10103498, Received: 20 August 2018; Accepted: 26 September 2018; Published: 29 September 2018, www.mdpi.com/journal/sustainability.

[7] Alessandro Massaro, Vincenzo Maritati, Angelo Galiano, Data Mining Model Performance Of Sales Predictive Algorithms Based On Rapidminer Workflows, *International Journal of Computer Science & Information Technology (IJCSIT)* Pg. No: 39-56, Vol. 10, No 3, June 2018.

[8] Aditya Joshi, Nidhi Pandey, (Professor) Rashmi Chawla, Pratik Patil, Use of Data Mining Techniques to Improve the Effectiveness of Sales and Marketing, *IJCSMC,* Vol. 4, Issue. 4, April 2015, pg.81 – 87, ISSN 2320–088X.

[9] Anurag Bejju, Sales Analysis of E-Commerce Websites using Data Mining Techniques, *International Journal of Computer Applications* (0975 – 8887), Pg. No:36-40, Volume 133 – No.5, January 2016.

[10] Abhilash CB, Rohitaksha K, Shankar Biradar, A Comparative Analysis of Data sets using Machine Learning Techniques, *IEEE International Advance Computing Conference (IACC),* Pg. 24-29, 2014, No: 978-1-4799-2572-8/14/$31.00 c 2014 IEEE.

[11] Lipika Dey, Hardik Meisheri and Ishan Verma, Predictive Analytics with Structured and Unstructured data - A Deep Learning based Approach, December 2017 Vol.18 No.2, *IEEE Intelligent Informatics Bulletin.*

[12] Marko Bohanec, Mirjana Kljaji Borstnarb, Marko Robnik-Sikonja, Explaining machine learning models in sales predictions, Preprint submitted to Elsevier November 20, 2016, Authors' post-print version - dx.doi.org/10.1016/j.eswa.2016.11.010.

DOI: 10.5373/JARDCS/V11I11/20193195
ISSN 1943-023X

Received: 16 Sep 2019/Accepted: 16 Oct 2019