International Journal on Emerging Technologies 11(2): 223-230(2020)

ISSN No. (Print): 0975-8364
ISSN No. (Online): 2249-3255

Analysis of Location based Sales Data using Machine Learning Algorithms

G.S. Ramesh[1], T.V. Rajini Kanth[2] and D. Vasumathi[3]
[1]Assistant Professor, Department of Computer Science and Engineering,
VNR VJIET, Hyderabad (Telangana), India.
[2]Professor, Department of Computer Science and Engineering,
SNIST, Hyderabad (Telangana), India.
[3]Professor, Department of Computer Science and Engineering,
JNTUH-CEH, Hyderabad (Telangana), India.

(Corresponding author: G.S. Ramesh)
(Received 18 December 2019, Revised 03 February 2020, Accepted 10 February 2020)
(Published by Research Trend, Website: www.researchtrend.net)

ABSTRACT: The sales data is exponentially increasing day by day and is very difficult to derive effective conclusions based on locations particularly if the data set consists of categorical and numerical attributes. It is equally important for the companies to retain the customers, enhance sales and build strategies to be upfront in the market. The major challenge is how to find sales importance based on location. Sales analytics based on geographical locations will be very useful to find insights, trends and helps to forecast the future sales. In this paper, initially the data set was pre-processed and suitable features selection was made using techniques namely Boruta & Variable Importance from Machine Learning Algorithms. Apart from thathybrid techniques were applied i.e. Cluster based Classification approaches were applied and comparison was made among various classifiers namely RBF Network, SMO, Random Tree, Naïve Bayesian Tree, etc. The classifiers Adaboost with J48, SMO and Random tree proved to be good when compared to all other classifiers. Location based analysis was also made by constructing maps.

Keywords: Sales Data, Hybrid Techniques, Cluster based classification approach, Location, Map.

I. INTRODUCTION

The studies show that it is difficult to find suitable factors that influence sales in terms of Geographic and Demographic domains so Geospatial Analytics plays major role in this aspect. The Sales Jan 2009 csv data set consists of "sanitized" trading records 30 days of the first month. All of the 998 records can be synopsized without any difficult and refined by records, country, city, date, payment and demography. We need to make a note of that these transactions are geocoded beforehand. Therefore we can utilize the already present latitude/longitude in the data available. The analysis of Sales data is vital way for companies in order to maximize the sales apart from meeting the customer needs in the competitive world. It is also used for Building strategies to retain customers, enhance sales and to increase new customer base apart from retaining the position in the upcoming competition from neighbours and new comers in the market.

The habit of examine of trading data produces accurate understanding from trading information and set benchmarks, predict upcoming trades. Examination must concentrate on increasing trades and innovate a new technique to increase trading performance in near future as well as for years to come.

A regular sales examination gives insights to organization regarding their performance where to improve and also their strengths. Each organization has its own targets, the employees of the organization must reach the targets. It is useful to estimate the position of the organization among other competitors with respect to sales and design a methodology to achieve the benchmark targets.

II. LITERATURE SURVEY

Joshi *et al.*, (2019) in their paper titled "Customer Centric Sales Analysis and Prediction" stated that to estimate the trades a lot of algorithms are like apriori, FP growth are used. They told that the money likely to be used by the consumers can be estimated with the help of other algorithms and differences among the algorithms are stated [1]. Kadam *et al.*, [2] stated that large shopping malls stores the trading information of each and every item to estimate upcoming needs and to manage goods stock. The information is analyzed to discover regular purchase patterns, deviations and also utilized for predicting upcoming volumes of trades by using random forests and multiple linear regression models.

Sastry *et al.*, (2013) in their paper titled "Analysis & Prediction of Sales Data in SAP-ERP System Using Clustering Algorithms" stated that the need for steel commodities is periodic and bank on lot of aspects such as consumer profile, amount, concessions charges levied. In this paper, they evaluated trades information by using clustering techniques such as K-Means and EM [3]. They discovered a lot of thought-provoking outlines which are helpful in increasing trades earnings and accomplishing greater trade numbers. Their study confirmed that partition methods like K-Means and EM algorithms are well-matched to analyze their sales data in contrast to Density based methods like DBSCAN and OPTICS or Hierarchical methods like COBWEB.

Wei and Peng (2013 stated that importance of projected investigation in this research article is to use the C4.5 process of information extraction in item- trade structure. The new simple probe for trade information is substituted with the actual aspects prompting for goods trades. The core study effort in this article contains the selection process in information extraction, elucidation and assessment of outcomes [4].

Beheshti-Kashi *et al.*, (2015) stated that this review article offers up-to-date techniques in the trades predicting investigation with an emphasis on the manner and novel goods prediction. The examination also analyses dissimilar tactics to the estimative cost of customer-produced data and search probes [5]. Wu *et al.*, (2019) stated that selling in advance is a promoting technique usually employed by online stores to surge trades by utilizing customer assessment ambiguity [6]. The aim is to search the whole consequence of permitting recompense on incomes from early trades, recognizing the situations where early selling with or without recompenses (or no early selling at all) are best. They compared systematically the returns of three early selling techniques: none, without recompense, and with recompense. The trading in advance and letting recompense is best for goods with a fairly minor return margin and lesser strategic marketplace scope, and that the additional returns can be substantial. The outcomes lead administrator in choosing the right early trading strategy. To enable this they, graphically display, established on the two dimensions of steady return margin and tactical market scope, under what circumstances the different techniques are best.

Donassolo and de Matos (2014) stated that finding trends within sales and discovering the most important factors affecting sales are interesting issues. The research aim of this article is to reason and illustrate the rectangle method for sales analysis. The core of the rectangle method is to find product items that may be offered to shops which have not ordered them [7]. Johnson (2016) stated that online panel information gathering infuses a number of benefits that sales researchers may provide [8]. This paper describes said advantages, limitations and also gives a analysis of latest trades associated examples using the information gathered. Besides, key attentions in using online panel information are advanced. The aim of the paper is to give a means to the trade scholars in enhancing their practice of online panel information.

Johnson *et al.,* (2014) stated that trade scholars are using more and more multilevel-multisource (MLMS) strategy to respond for numerous important queries containing trade executives, salespeople and consumers [9]. MLMS investigation includes the gained knowledge and examination of information gathered from many different origins relating to manifold stratified stages then displays many number of openings and contests for trade scholars to think about. Banking on this analysis, the writers propose a lot of impoverished sectors of investigation in which MLMS methods perhaps useful to more understanding of the vigorous situations that characterize trade examination. Zhi- Fan *et al.*, (2017) stated that online analyses offer customers with lot of data that might decrease their ambiguity concerns about procurements. The article proposes, a new technique which chains the Bass/Norton technique and belief investigation during employing past trade information along with online analysis of information by considering goods trade prediction. A belief examination technique, the Naive Bayes process, is applied to identify the belief directory from the gratification of every online analysis and combine it within the imitation coefficient of the Bass/Norton technique to develop the predicting efficiency. The calculated grades demonstrate that the blend of the Bass/Norton technique and belief investigation has greater predicting exactness than the standard Bass/Norton technique as well as various alternative trade prediction models [10].

Alfiah *et al.*, (2018) stated that apriori process as the root, that there are approaches of juxtaposition guidelines along with CRISP-DM [11]. Here the structure is able identify much o fthe goods of concern by consumers by employing data mining processes on each purchase information. Outcomes of data mining techniques to estimate trades trend in the direction of a trade goods, with this trades movement authority squad can study by revealing which goods trades seek sudden increase and which halt or decline. Where the successful goods of estimated number prediction that has a base worth above the verge base worth that has been determined. Pavlyshenko (2019) stated that the basic aim of their article is to deal with fundamental methods and examining different situation of applying machine learning techniques for trades prediction [12]. The results of machine-learning conclusions are taken into account. These results may be applied in order to identify trades forecast as soon as there is a minor chunk of past information for exact trade time sequence in the situation as soon as fresh goods released or else mart is opened. The outcomes show that applying pile up methods, we can increase the efficiency of forecasting prototypes for sales trades sequence prediction.

Mentzer and Moon (2004) stated that consolidating 25 years of trade's prediction handling study with above 400 organizations. Their process of study contains two main reviews of organizations trades prediction methods, a two-year, extensive research of trades prediction handling methods of 20 large-scale organizations, and continuing research of how to use the outcomes from the two-year research for directing trade prediction reviews of other organizations. It gave complete analysis of the methods and applications of trade prediction investigation [13].

Cerqueira *et al.*, (2018) stated that prediction is a vital job across quite a lot of areas. Its concluded attention is connected to the ambiguity and manifold evolving structure of time sequence [14]. Prediction techniques are normally intended to handle with temporal dependencies amongst research conclusion, but it is broadly acknowledged that nothing is commonly used. So, a typical answer to these jobs is to merge the conclusions of a varied set of predictions. In this paper they presented a method on arbitrating, in which numbers of prediction prototypes are vigorously pooled to achieve predictions. Outcomes from wide experimental trials give indication of the process effectiveness comparative to state of the art methods. The projected process is available publicly as a software bundle.

Tyralis and Papacharalampous (2017) stated that Time sequence prediction by applying machine learning techniques has increased demand in recent years [15]. Random forest is a machine learning technique applied on time sequence prediction; nevertheless, majority of its prediction features have remained uncharted. In this paper they concentrated on measuring the efficiency of random forests in one-step prediction by means of two big datasets of little time sequence with the focus to propose a best set of forecasting variables. The maximum forecasting efficiency RF is noted as soon as applying a less amount of relatively new diminishing forecasting variables. These results are helpful in appropriate forthcoming applications, with the view to obtain greater forecasting correctness.

III. PROPOSED METHODOLOGY

The Sales data considered for analysis is subjected to Pre-processing i.e. removal of Noise, filling up of Missing values with mean/mode. Identification of Outliers was done using Grubb's test, which is used to find a single outlier from a normally distributed data. Selecting relevant characteristics is the procedure of selecting probable items/features that are helpful in forecasting. Whenever developing forecasting techniques, always determine what characteristics are crucial and which is very helpful. Selection of Attributes/characteristics was done using the Feature selection methods namely Boruta & Variable Importance from Machine Learning Algorithms. Boruta is a feature ranking and selection algorithm based on random forests algorithm. Then applied Explorative data analytics (EDA) similar to Box Plots, Scatter Plots, Histogram etc. for further analysis. After that the refined pre-processed data set was subjected to k-Means Cluster algorithm with k = 5 clusters. The subsequent grouped information set i.e. Clustered data set was subjected to various Classifiers namely RBF Network, SMO, Random Tree, Naïve Bayesian Tree, Naïve Bayes, AdaBoost with Decision Stump and AdaBoost with J48 tree under the concept of Hybrid Data Mining approach. The results thus obtained were compared based on performance evaluation metrics namely Kappa statistic, F-Measure etc and concluded. The Proposed approach shown in Fig. 7 is proved to be useful in terms of Geo spatial Analytics apart from that the hybridized techniques i.e. clustered based classifiers improves their performance shown with Adaboost with J48 classifiers.

IV. RESULTS

Explorative data analysis was done on the Sales Data Set of Sanitized Products were analyzed by means of Scatter plot, Box plot, Bar chart/Histogram, Time sequence plot etc. The Histogram of payment Type, Product sales and Price were shown in the following graphs represented as Fig. 1 (a), Fig.1 (b) and (c). The payment type in Fig. 1 (a) is indicating that most of the customers across globe are using Visa card > 500 followed by Master Card which then followed by Amex and Diners. It is observed that most of the customers are using Visa Card than any other credit cards in payments. The Product Type sales in Fig. 1 (b) is indicating that Product-1 sales are > 800 and is very high when compared to others products types. Fig. 1 (c) shows that the sales of product Price Rs1200 is high followed by price Rs. 3600. Fig. 1 (d) shows Outliers Plot of Latitude with outlier value -41.47.

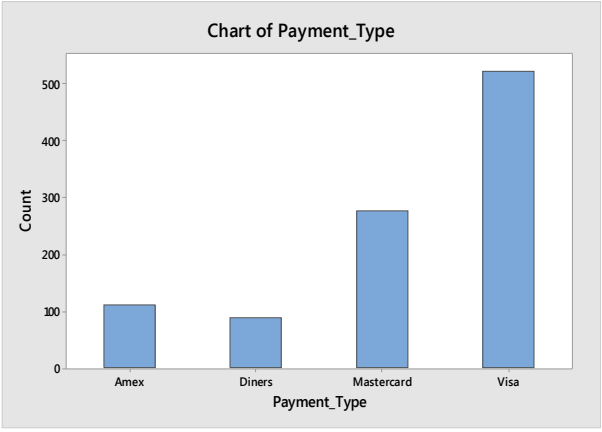

Fig. 1 (a) Payment Type.

Table 1.

Card type	Number of records
Amex	110
Diners	89
Mastercard	277
Visa	522

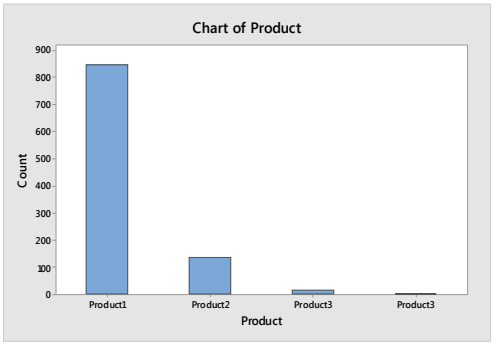

Fig. 1 (b) Product Type sales.

Table 2.

Product type	Number of records
Product1	847
Product2	136
Product3	15

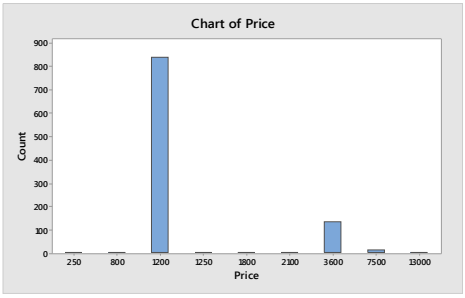

Fig. 1 (c) Price Chart.

Table 3.

Price chart	Number of records
250	1
800	1
1200	841
1250	1
1800	1
2100	1
3600	136
7500	15
13000	1

Fig. 2 (a) shows the 3D plot of Price between Latitude and Longitude i.e. how price variation between geographic Locations. Most of the products Price values varies from Rs 1000 to Rs 8000, between Latitude values varies between -100 to 180 and Longitude values varies from -40 to 45. Fig. 2 (b) shows the Country wise sale Price and it reveals that highest sales have taken place in United States followed by United Kingdom and Canada. Second highest level sales are across Switzerland, Ireland followed by Australia. Next highest level of sales is across France, Germany, Italy and Netherland countries. Fig. 2 (c) Map showing Country wise Sales on World Map and it shows that most of the customers are from European countries, second highest number of countries is from United States and followed by few countries from Asia.

The Fig. 3 shows how many distinct types of cards are used by the countries and it indicates 32 countries are using only one type of card whereas 9 countries NewZealand, United Kingdom, Australia, Germany, Canada, Italy, France, Switzerland and United States use all 4 types of cards. Eight countries use 2 types of cards and 7 countries use 3 types of cards out of 56 Fig. 4 shows the Box plot of the Data set 12 variables almost no outliers found except in Latitude and Longitude co-ordinates.

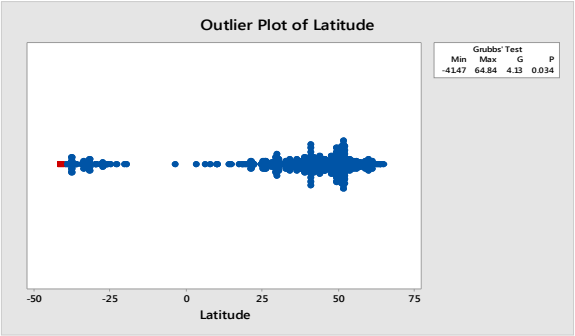

Fig. 1 (d) Outliers Plot of Latitude.

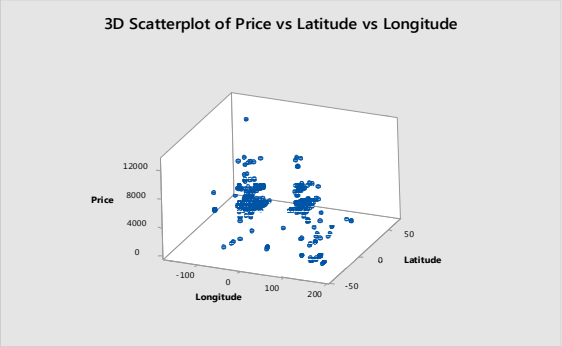

Fig. 2 (a) 3D Plot Price Vs Lat & Lon.

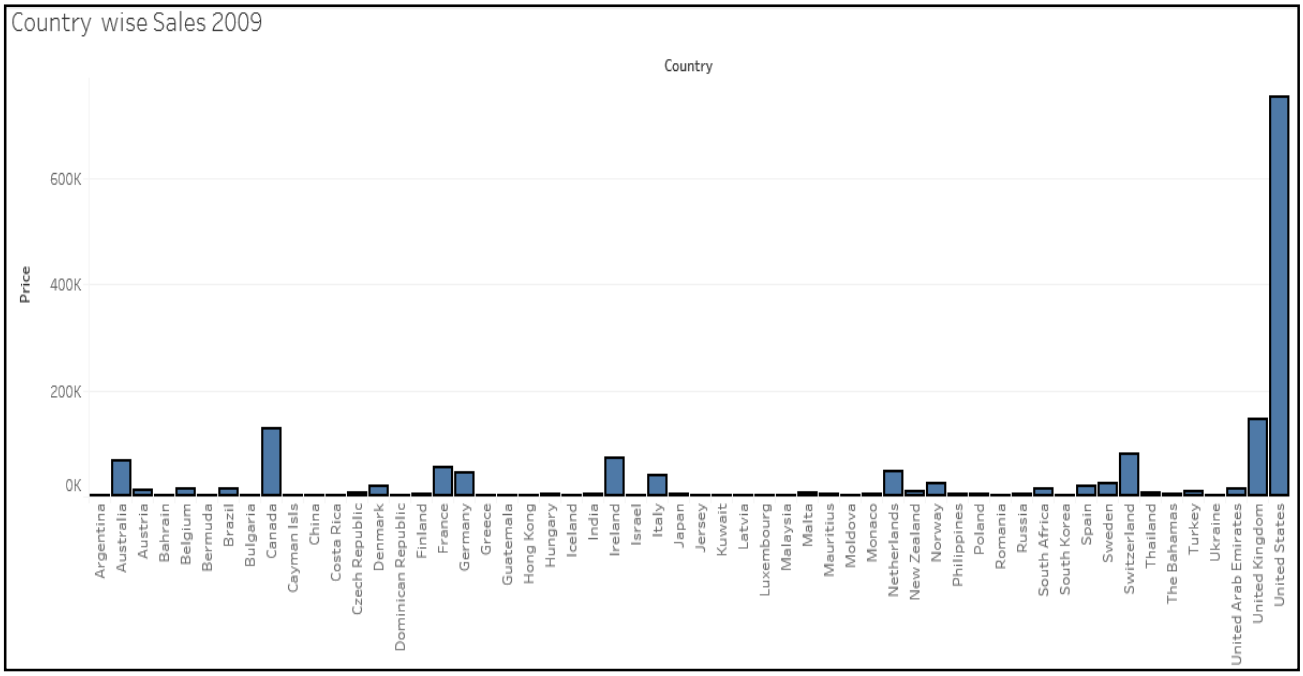

Fig. 2 (b) Country wise Products Sales.

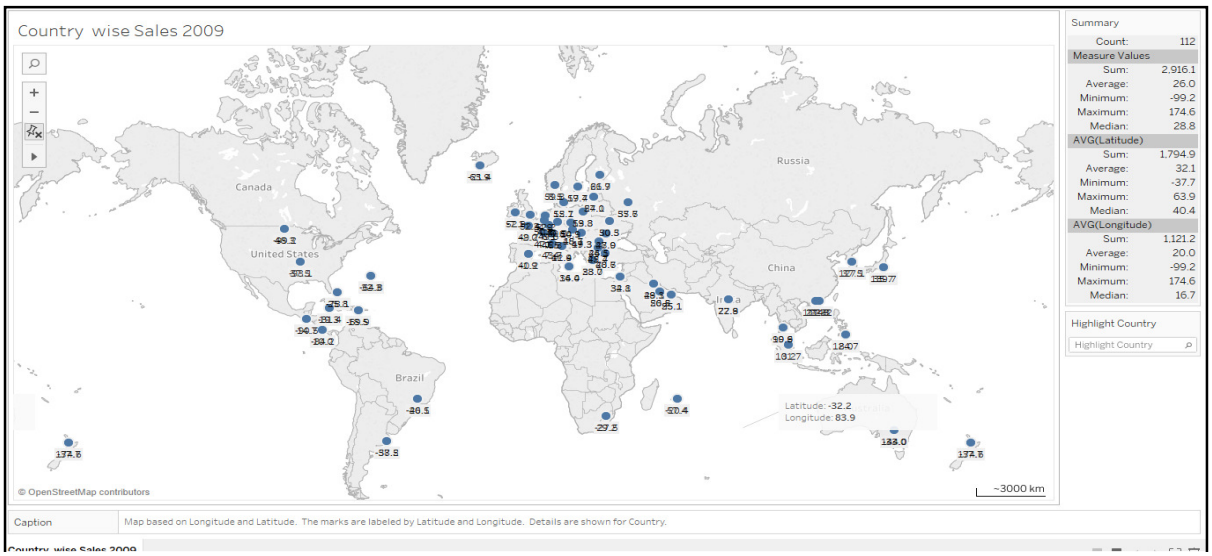

Fig. 2 (c) Map showing Country wise Sales.

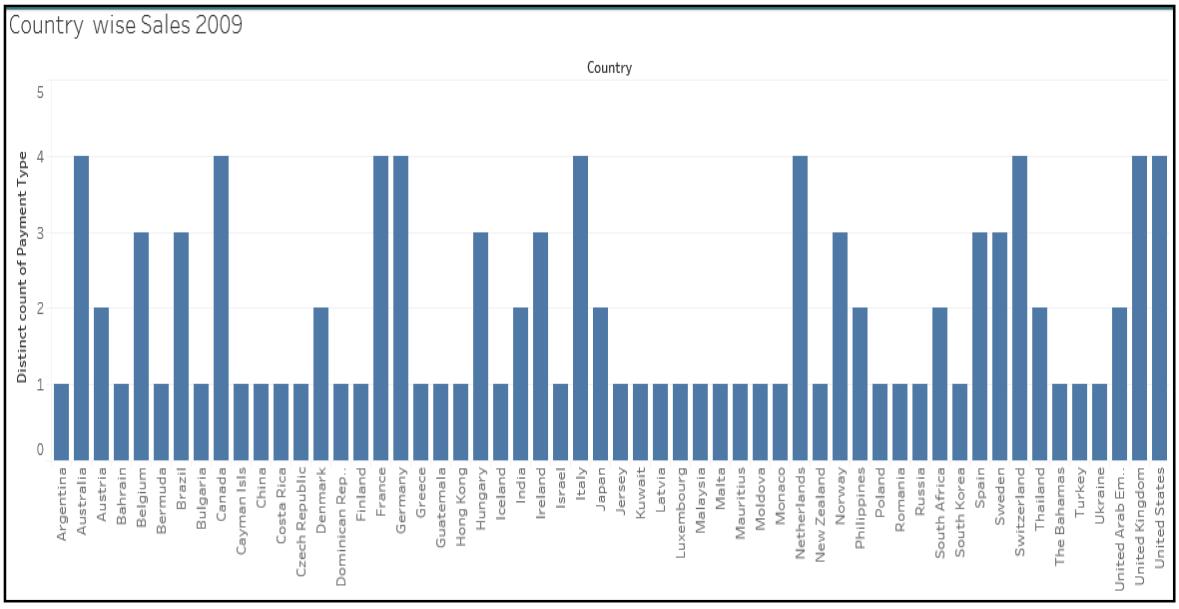

Fig. 3. Country wise Sales using Distinct Payment Cards Types.

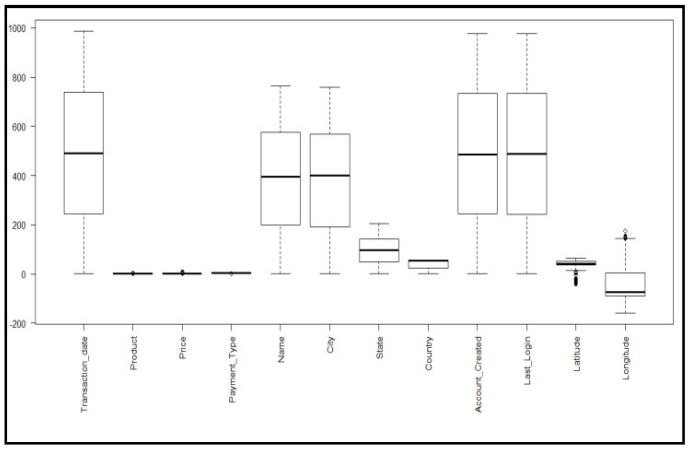

Fig. 4. The Box plot of 12 variables of the Data Set.

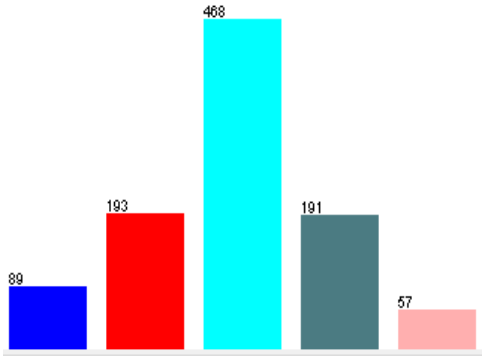

Fig. 5. Clusters Graph.

The Cluster Table-4 indicates the 5 Clusters with Names cluster (First)-0(89), cluster (second)-1(193), cluster (third)-2(468), cluster (Fourth)-3(191) and cluster (fifth)-4(57). Cluster-0 & Cluster-4 has Product Type -2 sales with sales values Rs.3600/- where as Cluster(second)-1, Cluster(third)-2 and Cluster(fourth)-3 has Product Type -1 sales with sales values Rs.1200/-. Cluster-3 & Cluster-4 has Payment_Type Diners where as Cluster-0, Cluster-1 & Cluster-2 has Payment_Type Master Card. Cluster-0 & Cluster-2 has sales in US where as Cluster-1 & Cluster-3 has sales near By Countries UK& Ireland and Cluster-4 has sales in Switzerland. The clustered Data set has instances with 89 (9%) for Cluster-0, 193 (19%) for Cluster-1, 468 (47%) for Cluster-2, 191 (19%) for Cluster-3 and 57 (6%) for Cluster-4 shown in Fig. 5.

It is concluded from Table4 that alone passive or active systems are not appropriate and sustainable due to increasing energy demand trend in space heating/cooling. It forces us to adopt suitable hybrid systems according to tailor made situations.

Table 4: K-Means Cluster – 5 Clusters Values.

Attribute	Full Data (998)	Cluster-0 (89)	Cluster-1 (193)	Cluster-2 (468)	Cluster-3 (191)	Cluster-4 (57)
Product	1.1663	2.1348	1.0363	1	1.0262	1.9298
Price	1200	3600	1200	1200	1200	3600
Payment Type	2.011	2.1236	2.2487	2.0449	1.7225	1.7193
Name	Sarah	Michael	Lisa	Sarah	Stephanie	Family
City	London	Toronto	London	Calgary	Den Haag	Lausanne
State	England	CA	England	CA	Dublin	Zurich
Country	US	US	UK	US	Ireland	Switzerland
Latitude	39.0157	39.3641	29.7374	38.4987	51.2071	33.2809
Longitude	-41.3378	-82.8449	36.8969	-92.6885	3.5202	29.8732

Table 5: Performances Comparison of 9 classifiers.

Classifier Techniques	RBF Network	SMO	Naive Baye's	Random Forest	Random Tree	J48	Ada Boost With Decision Stump	NB Tree	Ada Boost With J48
% of Correctly Classified Instances	97.495	100	95.992	99.6848	100	97.1944	65.7315	98.6974	100
Kappa Statistics	0.9639	1	0.9423	0.9952	1	0.9595	0.4782	0.9812	1
Precision	0.975	1	0.962	0.997	1	0.972	0.482	0.987	1
Recall	0.975	1	0.96	0.997	1	0.972	0.657	0.987	1
F-Measure	0.992	1	0.96	0.997	1	0.971	0.545	0.987	1
Mean Absolute Error	0.0185	0.24	0.0169	0.0024	0	0.0172	0.2664	0.0065	0
Relative Absolute Error	6.6629	86.2837	6.0617	0.8949	0	6.1835	95.7754	2.3214	0
Root Mean Square Error	0.096	0.3162	0.1146	0.0325	0	0.0927	0.3411	0.0674	0
Time taken to create the model in Sec	2.51	19.27	0.02	6.69	0.02	0.03	0.03	9.63	0.22

The Table 5 shows the performances of 9 Classifiers based on 9 parameters namely % of Properly categorized Instances, Kappa Statistic, Precision, Recall, F-Measure, Mean Absolute error, Relative Absolute error, Root Mean Square error and period of Time required to create the prototype in Sec. It was observed that out of all 9 classifiers SMO, Random Tree and Ada Boost with J48 classifier proved to be best classifiers.

The Classifiers Random Tree out performed across all the Classifiers next followed by Ada Boost with J48 classifier with more time to build the model as it is Hybrid Model. So Random Tree Classifier is best suitable than all other classifiers in terms of performance for the Sales data set. The Price can be calculated at a particular location using the Linear Regression Equation given by Eqn. (1). The result is shown in Fig. 6.

Price = 1734.5 – 2.90 Latitude – 0.304 Longitude (1)

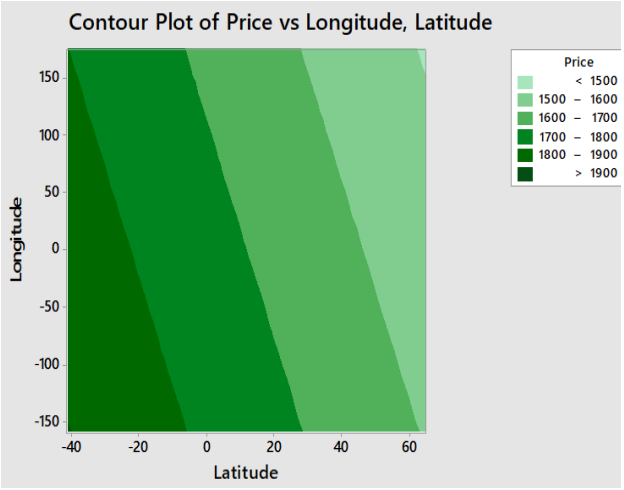

Fig. 6. Location Based Contour Plot for Price.

Table 6: Location Based Price Prediction.

Latitude	Longitude	Price Prediction
-42	-160	1905.05
5	-120	1756.44
23	-100	1698.11
58	22	1559.39
65	176	1492.22

Table 6 shows the Prediction Values of Price based on location using Eqn. (1) when Unknown values of Latitude and Longitude values are given. Most of studies reveals that they mostly depended on Association algorithms like Apriori, FP Growth and simple K-means clustering and EM algorithms where as our Proposed methodology was extended further over clustering techniques like Hybrid Techniques and also drawn Maps for effective Location based Analysis for sales prediction.

V. CONCLUSION

Most of the customers across globe are using Visa card followed by Master Card and followed by Amex and Diners for payments. It is observed that most of the customers use Visa Card than any other credit cards in payments. The Product-1 sales are very high when compared to others Product Types. The sales of product Price Rs 1200 is high followed by price Rs 3600. Most of the products Price values vary from Rs1000 to Rs 8000, between Latitude values -100 to 180 and Longitude values -40 to 45. The Country wise sales Price are highest in United States followed by United Kingdom and Canada. Second highest level sales are across Switzerland, Ireland followed by Australia. Next highest level of sales is across France, Germany, Italy and Netherland countries. Country wise Sales on World Map indicate that most of the customers are from European countries followed by United States and then followed by few countries in Asia. Out of 56 countries 32 countries are using only one type of credit card with low sales and whereas 9 countries are using all 4 types of countries with highest sales. Random Tree Classifier is the best suitable algorithm than all other classifiers in terms of performance for the Sales data set. The same result can be obtained using Adaboost classifier with J48 in which it takes higher time to build model

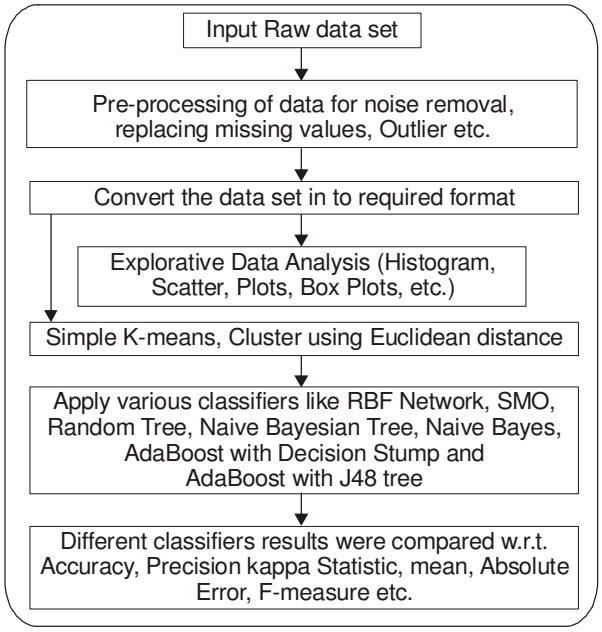

Fig. 7. The Proposed Methodology.

VI. FUTURE SCOPE

The proposed methodology can be combined with Deep Learning Techniques for further improvement of the trade predication.

ACKNOWLEDGEMENTS

Special Thanks to Spatial Key Support organization for the data set Sales Jan 2009

REFERENCES

[1]. Joshi, S., Rao, L. S., & Ida, B. (2019). Seraphim, Customer Centric Sales Analysis and Prediction. *International Journal of Engineering and Advanced Technology (IJEAT), 8*(4), 1749-1753.

[2]. Kadam, H., Shevade, R., Ketkar, D., & Rajguru, S. (2018). A Forecast for Big Mart Sales Based on Random Forests and Multiple Linear Regression. *International Journal of Engineering Development and Research, 6*(4), 41-42.

[3]. Sastry, S. H., Babu, P., & Prasada, M. S. (2013). Analysis & Prediction of Sales Data in SAP-ERP System using Clustering Algorithms. *International Journal of Computational Science and Information Technology (IJCSITY), 1*(4), 95-109.

[4]. Wei, T., & Peng, G. (2013). Research on Retail Sales Management System Based on Data Mining Technology. In *Joint International Conference on Pervasive Computing and the Networked World*, 586-592. Springer, Cham.

[5]. Beheshti-Kashi, S., Karimi, H. R., Thoben, K. D., Lütjen, M., & Teucke, M. (2015). A survey on retail sales forecasting and prediction in fashion markets. *Systems Science & Control Engineering*, *3*(1), 154-161.

[6]. Wu, M., Teunter, R. H., & Zhu, S. X. (2019). Online marketing: When to offer a refund for advanced sales. *International Journal of Research in Marketing*, *36*(3), 471-491.

[7]. Donassolo, P. H., & de Matos, C. A. (2014). The predictors of sales performance: a study with wholesale sellers. *Revista Brasileira de Gestão de Negócios-RBGN*, *16*(52), 448-465.
[8]. Johnson, J. S. (2016). Improving online panel data usage in sales research. *Journal of Personal Selling & Sales Management*, *36*(1), 74-85.
[9]. Johnson, J. S., Friend, S. B., & Horn, B. J. (2014). Levels of analysis and sources of data in sales research: a multilevel-multisource review. *Journal of Personal Selling & Sales Management*, *34*(1), 70-86.
[10]. Fan, Z. P., Che, Y. J., & Chen, Z. Y. (2017). Product sales forecasting using online reviews and historical sales data: A method combining the Bass model and sentiment analysis. *Journal of Business Research*, *74*, 90-100.
[11]. Alfiah, F., Pandhito, B. W., Sunarni, A. T., Muharam, D., & Matusin, P. R. (2018). Data Mining Systems to Determine Sales Trends and Quantity Forecast Using Association Rule and CRISP-DM Method. *International Journal of Engineering and Techniques*, *4*(1), 186-192.
[12]. Pavlyshenko, B. M. (2019). Machine-learning models for sales time series forecasting. *Data*, *4*(1), 1-11.
[13]. Mentzer, J. T., & Moon, M. A. (2004). *Sales forecasting management: a demand management approach*. Sage Publications.
[14]. Cerqueira, V., Torgo, L., Pinto, F., & Soares, C. (2018). Arbitrage of forecasting experts. *Machine Learning*, *108*(6), 913-944.
[15]. Tyralis, H., & Papacharalampous, G. (2017). Variable selection in time series forecasting using random forests. *Algorithms*, *10*(4), 1-25.
[16]. Ramesh, G. S., RajiniKanth, T. V., & Vasumathi, D. (2019) Effective Analysis of Sales Data Set Using Advanced Classifier Techniques. *Journal of Advance Research in Dynamical & Control Systems*, *11*(11), 260-266.

How to cite this article: Ramesh, G. S., Rajini Kanth, T. V. and Vasumathi, D. (2020). Analysis of Location based Sales Data using Machine Learning Algorithms. *International Journal on Emerging Technologies*, *11*(2): 223–230.