Intelligent Sales Prediction Using Machine Learning Techniques

Sunitha Cheriyan
IT Department
Higher College of Technology
Muscat, Sultanate of Oman
sunitha.cheriyan@hct.edu.om

Shaniba Ibrahim
IT Department
Higher College of Technology
Muscat, Sultanate of Oman
shaniba.ibrahim@hct.edu.om

Saju Mohanan
IT Department
Higher College of Technology
Muscat, Sultanate of Oman
saju.mohanan@hct.edu.om

Susan Treesa
IT Department
Higher College of Technology
Muscat, Sultanate of Oman
susan@hct.edu.om

***Abstract*— Intelligent Decision Analytical System requires integration of decision analysis and predictions. Most of the business organizations heavily depend on a knowledge base and demand prediction of sales trends. The accuracy in sales forecast provides a big impact in business. Data mining techniques are very effective tools in extracting hidden knowledge from an enormous dataset to enhance accuracy and efficiency of forecasting. The detailed study and analysis of comprehensible predictive models to improve future sales predictions are carried out in this research. Traditional forecast systems are difficult to deal with the big data and accuracy of sales forecasting. These issues could be overcome by using various data mining techniques. In this paper, we briefly analyzed the concept of sales data and sales forecast. The various techniques and measures for sales predictions are described in the later part of the research work. On the basis of a performance evaluation, a best suited predictive model is suggested for the sales trend forecast. The results are summarized in terms of reliability and accuracy of efficient techniques taken for prediction and forecasting. The studies found that the best fit model is Gradient Boost Algorithm, which shows maximum accuracy in forecasting and future sales prediction.**

Keywords*— *Data mining techniques, Machine Learning Algorithms, Prediction, Reliability, Sales forecasting

I. INTRODUCTION

One of the major objectives of this research work is to find out the reliable sales trend prediction mechanism which is implemented by using data mining techniques to achieve the best possible revenue. Today's business handles huge repository of data. The volume of data is expected to grow further in an exponential manner. The measures are mandatory in order to accommodate process speed of transaction and to enhance the expected growth in data volume and customer behavior. The E-commerce industry is badly in need of new data mining techniques and intelligent prediction model of sales trends with highest possible level of accuracy and reliability. Sales forecasting gives insight into how a company should manage its workforce, cash flow and resources. It is an important prerequisite for enterprise planning and decision making. It allows companies to plan their business strategies effectively.

Accurate predictions allow the organization to improve market growth with higher level of revenue generation. Data mining techniques are very effective in tuning huge volume of data into useful information for cost prediction and sales forecast, it is the basic of sound budgeting [1]. At the organizational level, forecasts of sales are essential inputs to many decision making activities in various functional areas such as operations, marketing, sales, production and finance. In order to serve an organization's internal resources effectively, predictive sales data is important for businesses when looking for acquiring investment capital. The studies proceed with a new perspective that focuses on how to choose an appropriate approach to forecast sales with high degree of precision. Initial dataset considered in this research had a large number of entries, but the final dataset used for analysis having much smaller size compared to the original due to the riddance of non-usable data, redundant entries and irrelevant sales data.

The data mining techniques and predictions methods are discussed in Section I. The review of various literatures about sales forecasts are stated in Section II. In Section III, data tuning process and predictions are highlighted with visual representation of generated results. The predictive analytics and methodology on sales price also discussed. The performance evaluations of various prediction algorithms using machine learning approaches are stated. Finally, the result is analyzed and concluded by summarizing the research findings and future scope.

II. RELATED WORK

In order to be competent enough and to generate higher revenue, business organizations are constantly in search of a better model or technique for data mining and maintenance of critical data [2]. Business industry faces severe challenges to identify an accurate data mining technique and effective predication strategy [3] due to the exponential growth of huge volume of data used in e-commerce transactions. Sales data analysis faces lot of issues and major aspects of sales functions are identification of product attribute, price fixation, net sales realization and launch of new product. Various prediction methods, sales forecasting strategies and Expectation Maximization (EM) algorithm are discussed in [4].

A comparative study on data tuning and various clustering algorithms on sales data is clearly explained in [5]. As analyzed in [6], classification of data is very important in decision making. Clustering techniques are very useful in discovering distribution patterns and clustering algorithms employ a distance metric based similarity measures [7]. In an appropriate data mining techniques information from a bulky data set can be transformed into a reasonable format and can be done by using supervised and unsupervised learning [8]. With an appropriate sales prediction technique, effective business decision making can be done. The concepts and algorithms are handled in [9]. As suggested by Korolev and Ruegg, the prediction error can be reduced with the implementation of XGBoost and additional support of SigOpt Bayesian Optimization method [10]. Sales forecasting can be done using different datamining techniques where predicting sales on any given day at any

store can be carried out, the detailed analysis and procedures are shown in [11].

In this project we have performed sales forecasting for stores using different data mining techniques. The task involved predicting the sales on any given day at any store, in order to familiarize ourselves with the task we have studied previously.

III. RESEARCH METHODOLOGY

The main purpose of this research is to evaluate and analyze the use of data mining techniques for sales forecasting, to produce models which are comprehensive and reliable.

A. Data Collection and Preparation

The dataset used for this research is based on an e-fashion store, for the three consecutive years of sales data. To predict the sales of the e-fashion store, past sales record for three years from 2015 to 2017 were collected. The database includes Category, City, Type of items and its description, number of items, Quantity, Quarter, Sales Revenue, Year, SKU description, Week, Year. The data consisted initially of large number of entries, but the final selected dataset had much smaller size compared to the original dataset due to removal of non-usable data, redundant entries and irrelevant data[12].

B. Exploratory Analysis

After data preprocessing, in order to clearly understand the nature of our data, an exploratory analysis was conducted[13]. The exploratory analysis consists of the steps as shown in the Figure 1.

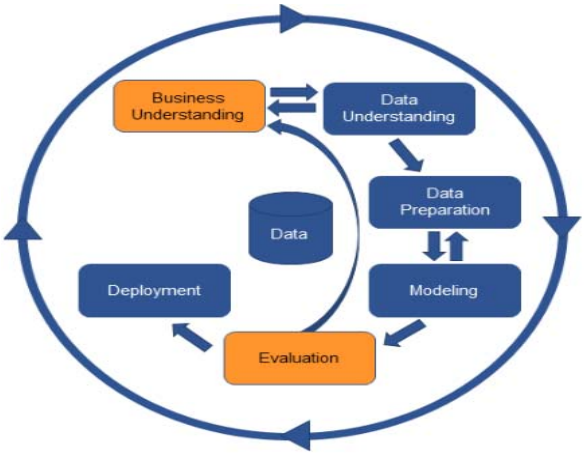

Figure 1. Stages of Data mining

The stages involved in the data mining model include data understanding, preparation, modelling, evaluation and deployment. The analysis of the collected sales data is shown in the Table I.

TABLE I. YEARLY SALES DATA

sales	Year of Data		
Quarter	2015	2016	2017
1	2,425,084	3,338,276	3,032,690
2	1,623,062	2,852,142	2,447,278
3	1,254,177	2,885,557	3,306,000
4	1,795,108	4,199,956	2,107,128

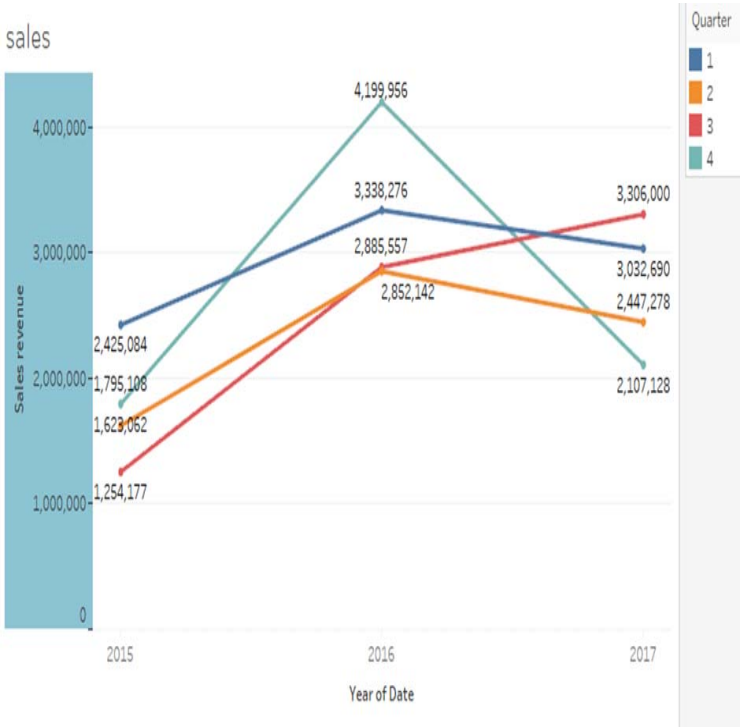

Figure 2. Yearly Sales Visualization

The Figure 2 shows the visualization about the amount of sales revenue generated in the years from 2015 to 2017 and in the quarters represented as 1,2,3,4 respectively. This shows that in quarter 4 of 2016 the sales revenue is high and in quarter 3 of 2015 shows a great decrease in the revenue generated.

C. Outlier detection

This process performs all necessary data preprocessing and model optimization. Outlier detection process can be used to deploy the model or as a starting point for further optimizations and helpful in showing generic information which is independent of the models. The main focus is on the quality of the data, especially the quality of each data attributes. Besides, these also consider discarding the data attributes that provide less value.

Data: the dataset after it has been transformed for modeling.

Correlations: a matrix showing the correlations between the attributes with a positive correlation on the sales revenue which is given in the Table II.

TABLE II. CORRELATION MATRIX

Attribute	Quantity	Sales Revenue	SKU Description	Week
Quantity	1	0.947	-0.294	-0.021
Sales Revenue	0.947	1	-0.247	-0.044
SKU Description	-0.294	-0.247	1	-0.269
Week	-0.021	-0.044	-0.269	1

D. Forecasting and Trends

The figure 3 shows the forecasting of the future sales from Quarter 3 of 2018 to Quarter 3 of 2021. The trend shows the sum of sales revenue for the dated Quarter. The blue color indicates the actual sales generated and red color indicates the estimated sales for the dated quarter showing a slight increase in the sales as shown in Figure 3.

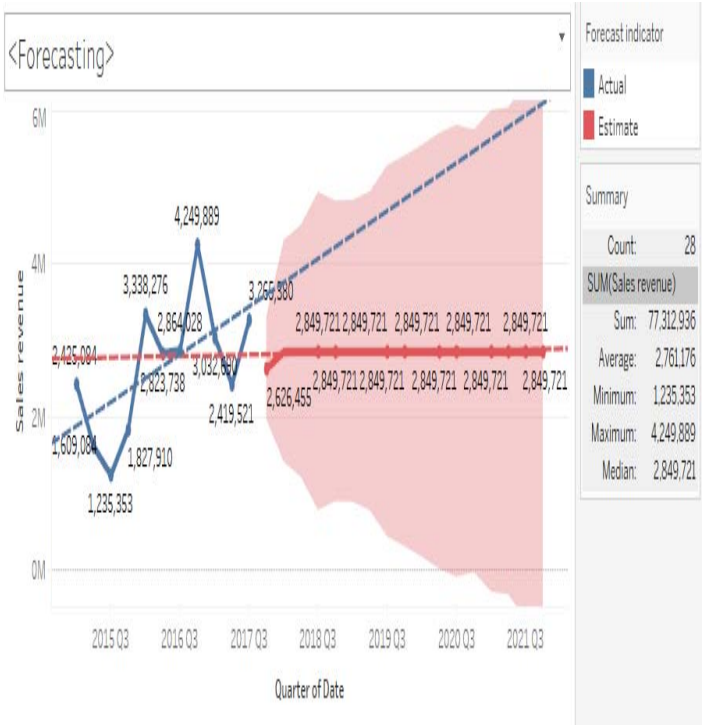

Figure 3. Forecast for five years

Further the trend is also generated by grouping the Sales Revenue and the sum of Quantity sold for quarter into Cluster 1 and Cluster 2 respectively. The blue color line indicates Cluster 1 and the orange color represents the Cluster 2. The result shows a trend line showing a slight dip in the Quantity sold and the Sales revenue in the Cluster 1 and 2 in the fourth Quarter as shown in the Figure 4.

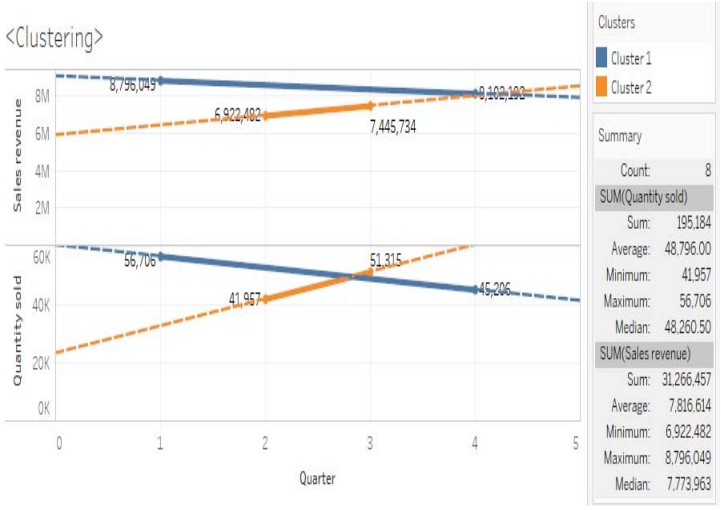

Figure 4. Trend Analysis using Clusters

The forecast is composed of a smoothed averaged adjusted for a linear trend. Then the forecast is also adjusted for seasonality. The Figure 5 shows the details about the model used for the trend analysis. The model shows a seasonal effect high in the month of January 2022 and low in the month of August 2022.

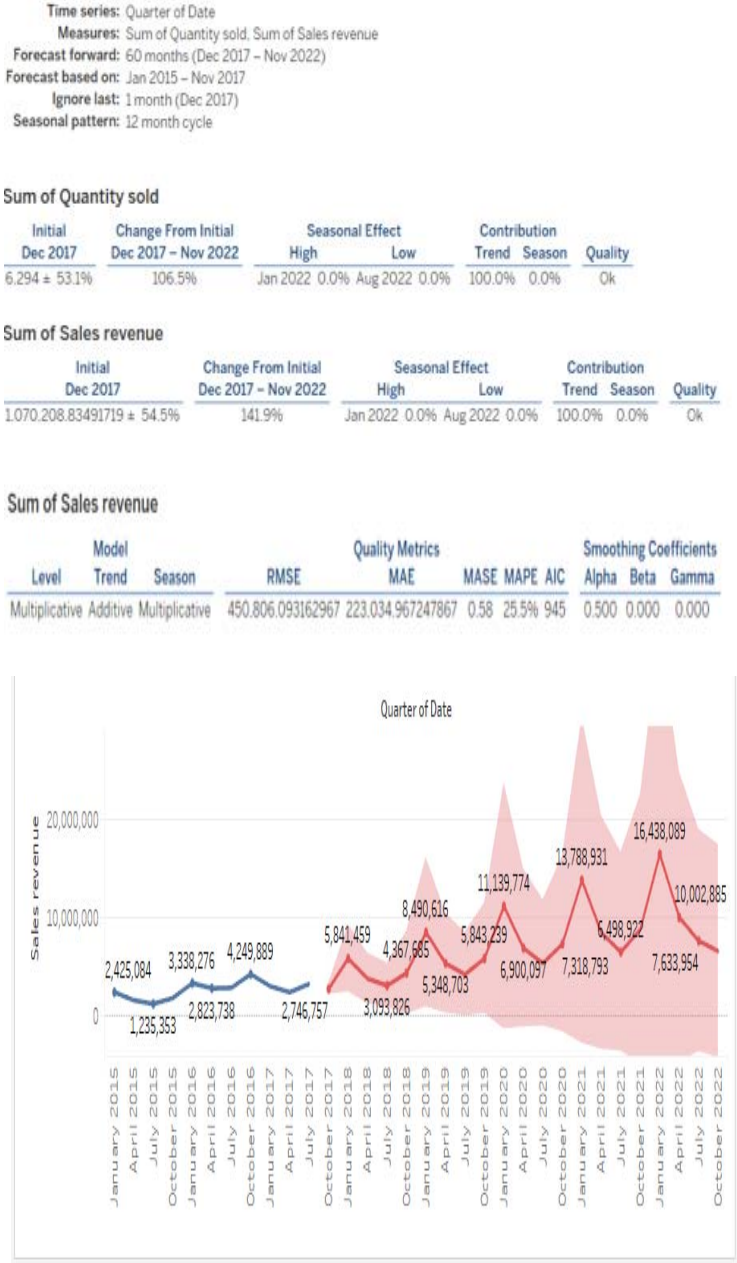

Figure 5. Trend Analysis using Exponential Smoothing

E. Prediction

Prediction deals with events occurring in the future. The use of Machine learning algorithms improves the intelligence of the system without manual intervention. “Machine Learning (ML) is used to optimize the performance criterion using sample data or the past experience” as defined by Ethem Alpaydin [14].

Machine learning techniques can be applied to all disciplines. Machine learning uses statistics to solve many classification and clustering problems. The ML algorithms are classified in three categories [15]. They are supervised, unsupervised and semi supervised. In this paper we discussed about three machine learning algorithms which can be applied to prediction, like Generalized Linear Model (GLM), Decision Tree (DT) and Gradient Boost Tree (GBT).

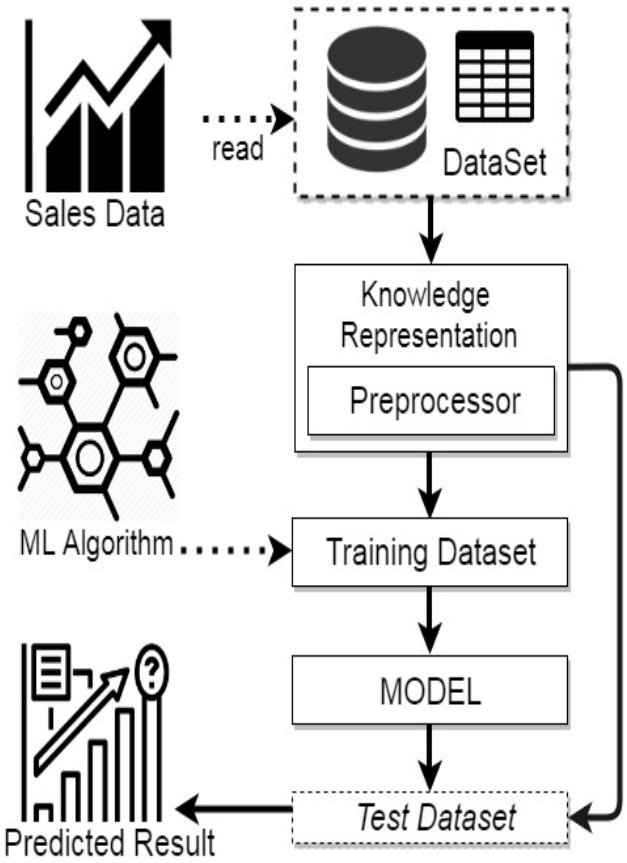

Figure 6. System Architecture

In this study, we implemented three machine learning algorithms on the training dataset and the models are tested for the performance. Based on the performance accuracy the best algorithm is chosen for the prediction.

1) Generalized Linear Model

Generalized linear model (GLM) refers to a large class of conventional linear regression model [16]. The focus is for a continuous response where the variable gives continuous categorical predictors [17]. One of the major components in a generalized linear model is a random component which is the probability distribution of the response variable (Yi);
A linear predictor is another important component, which can be represented as:

$$\eta i = \alpha + \beta 1Xi1 + \beta 2Xi1 + \cdots + \beta kXik \quad \text{---------- (1)}$$

A smooth and invertible linearizing link function g(.), which transforms the

$$\mu i = E(Yi) \quad \text{---------- (2)}$$

$$g(\mu i) = \eta(i) = \alpha + \beta 1Xi1 + \beta 2Xi2 + \cdots . + \beta kXik$$

---------- (3)

Generalized linear models are providing estimate of the regression coefficients and estimated asymptotic standard errors of the coefficients. Usually the dispersion parameter in GLS is fixed to a numeric value 1[18].

2) Decision Tree

Decision tree is a classifier referred as recursive partition of the instant space. It is a powerful form of multiple variable analyses and is a strong data mining tool. Its applications are found in various domains and this approach represents factors involved in achieving a predetermined goal and the corresponding factors to achieve the goal and the ways and means of implementation. [14] Let the objective can be denoted as (O) and (C_i) is the ways to follow and let (M_{ij}) the means of action corresponding to these ways, which can be noted by qi, (i= 1 …. n), which meets the relation.

$$\sum_{i=1}^{n} qi = 1, cu\ qi \geq 0 \quad \text{…….. (4)}$$

For the means of action (M_{ij}), the important coefficient (a_{ij}), includes set of weights, where the sum is equal to 1 for each way $a_{11}+a_{12}+\ldots..+a_{1m} = 1$,

$a_{21}+a_{22}+\ldots..+a_{2m} = 1$

$\ldots\ldots\ldots$

$a_{n1}+a_{n2}+\ldots\ldots..+a_{nm}= 1$

$$\sum aij = 1 \quad \text{…… (5)}$$

The Figure 7 is an example of decision tree model displaying the different items as a tree in each quarter and the root node is the year.

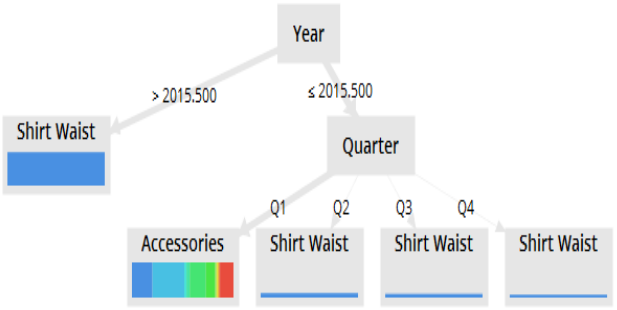

Figure 7. Decision Tree Model

3) Gradient Boosted Trees

Gradient boosting is a machine learning technique for regression and classification problem. This approach could ensemble learning method that combines large number of decision trees to produce final prediction model [10]. This model is built on a principle that a collection of weak learners combined together can produce a strong learner by using boosting process. GBT approach has a strong additive

training method, required for adding a new weak learner into the model, the weak learner is the decision tree [19].

Let F(x) is a full model after t-1 round and h(x) is the new tree, added to the model. F0= 0 ……. (6)

Ft (x) = Ft-1(x) +h(x) ..…….. (7)

Each new function is an attempt to correct the errors of the model built in previous rounds. Hence the new function(x) must be able to predict the residual Ft-1(x). The Figure 8 shows a sample of gradient boosted tree.

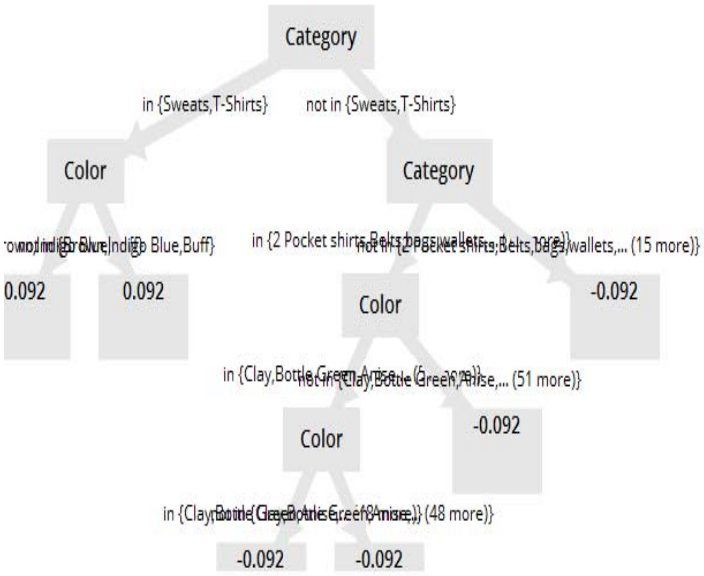

Figure 8 . Gradient Boosted Tree Model

IV. RESULT AND ANALYSIS

The performance of the classification algorithms is mostly focused on Classification accuracy, Accuracy in each class and confusion matrix which shows the number of predictions of each class which can be compared to the instances of each class. Root Mean Square Error, Mean Square Error, Absolute error are calculated and average of the error is shown in the output in the Table III as the Error Rate. This measure helps to identity whether the given prediction is wrong on average.

The comparative studies of the three algorithms based on the prediction performance are given in the Table 1 and the visualization in Figure 9. Based on the performance, it is understood that Gradient Boost Algorithm is showing 98% overall accuracy and the second stands Decision Tree Algorithms with nearly 71% overall accuracy and followed by Generalized Linear Model with 64% accuracy. Finally, it can be compared based on the empirical evaluation of the three chosen algorithm the best fit for the model is Gradient Boosted Tree. The classification accuracy rate can reach up to 100%, but in GBT model analyzed and shown in Table III, achieved approximately 98% of accuracy. If there are further improvement on the GBT implementation with the support of a strong data set along with models such as Grabit, Tobit as analyzed in [20], projects better accuracy rate.

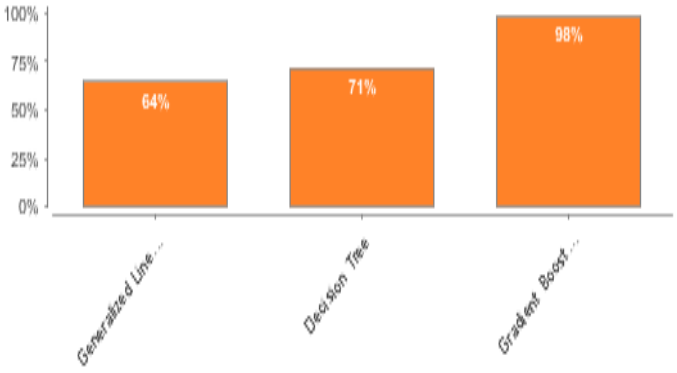

Figure 9. Comparison between the Models

TABLE III. PREFORMANCE SUMMARY

Model Name	Performance Summary of ML Algorithms				
	Accuracy Rate (%)	Error Rate	Precision	Recall	Kappa[1]
GLM	64	36	5.36	0	0
DT	71	29	11.24	15.61	0.501
GBT	98	2	50	50	0.962

The accuracy rate, error rate, precision, recall and Kappa values shown in Table III is a clear indication that Gradient Boosted Tree (GBT) stand out as pioneer model with the accuracy rate of 98 % and minimum error rate.

V. CONCLUSION

The researchers have concluded that an intelligent sales prediction system is required for business organizations to handle enormous volume of data. Business decisions are based on speed and accuracy of data processing techniques. Machine learning approaches highlighted in this research paper will be able to provide an effective mechanism in data tuning and decision making. In order to be competent in business, organizations are required to equip with modern approaches to accommodate different types of customer behavior by forecasting attractive sales turn over. In our studies, we used almost 85,000 records for the comparison of algorithms. Since the time of execution was huge and to manage such a large set of records are complex, some of the records were discarded, during the analysis phase. At the same time, fields and attributes, used in this analysis were insufficient for the further analysis. It was the major challenge we faced during the research. However, we had thoroughly weighed our works by implementing efficient ML techniques for prediction and forecasting. The current studies can be expedited by using Big Data as a tool for the predictive analytics in sales forecasting. The big data analysis and forecasting are measured as the vital fields in the modern business scenario.

REFERENCES

[1] Huang, Q., & Zhou, F. (2017, March). Research on retailer data clustering algorithm based on spark. In AIP Conference Proceedings (Vol. 1820, No. 1, p. 080022). AIP Publishing.

[2] Saylı, A., Ozturk, I., & Ustunel, M. (2016). Brand loyalty analysis system using K-Means algorithm. Journal of Engineering Technology and Applied Sciences, 1(3).

[3] Maingi, M. N. A Survey on the Clustering Algorithms in Sales Data Mining.

[4] Sastry, S. H., Babu, P., & Prasada, M. S. (2013). Analysis & Prediction of Sales Data in SAP-ERP System using Clustering Algorithms. arXiv preprint arXiv:1312.2678.

[5] Shrivastava, V., & Arya, N. (2012). A study of various clustering algorithms on retail sales data. Int. J. Comput. Commun. Netw, 1(2).

[6] Rajagopal, D. (2011). Customer data clustering using data mining technique. arXiv preprint arXiv:1112.2663.

[7] Tsai, C. F., Wu, H. C., & Tsai, C. W. (2002). A new data clustering approach for data mining in large databases. In Parallel Architectures, Algorithms and Networks, 2002. I-SPAN'02. Proceedings. International Symposium on (pp. 315-320). IEEE.

[8] Mann, A. K., & Kaur, N. (2013). Review paper on clustering techniques. Global Journal of Computer Science and Technology.

[9] Shah, N., Solanki, M., Tambe, A., & Dhangar, D. Sales Prediction Using Effective Mining Techniques.

[10] Korolev, M., & Ruegg, K. (2015). Gradient Boosted Trees to Predict Store Sales.

[11] Jain, A., Menon, M. N., & Chandra, S. Sales Forecasting for Retail Chains.

[12] Rey, T. D., Wells, C., & Kauhl, J. (2013). Using data mining in forecasting problems. In SAS Global Forum 2013: Data Mining and Text Analytics.

[13] Huang, W., Zhang, Q., Xu, W., Fu, H., Wang, M., & Liang, X. (2015). A Novel Trigger Model for Sales Prediction with Data Mining Techniques. Data Science Journal, 14.

[14] Ethem Alpaydin. (2004). Introduction to Machine Learning (Adaptive Computation and Machine Learning), The MIT Press.

[15] Lytvynenko, T. I. (2016). Problem of data analysis and forecasting using decision trees method.

[16] Lazăr, C., & Lazăr, M. (2015). Using the Method of Decision Trees in the Forecasting Activity. Petroleum-Gas University of Ploiesti Bulletin, Technical Series, 67(1).

[17] Flesch, B., Vatrapu, R., Mukkamala, R. R., & Hussain, A. (2015, October). Social set visualizer: A set theoretical approach to big social data analytics of real-world events. In Big Data (Big Data), 2015 IEEE International Conference on(pp. 2418-2427). IEEE.

[18] Asooja, K., Bordea, G., Vulcu, G., & Buitelaar, P. (2016). Forecasting Emerging Trends from Scientific Literature. In LREC.

[19] Stearns, B., Rangel, F., Rangel, F., de Faria, F. F., Oliveira, J., & Ramos, A. A. D. S. (2017). Scholar Performance Prediction using Boosted Regression Trees Techniques. In European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN). Citeseer.

[20] Sigrist, F., & Hirnschall, C. (2018). Gradient Tree-Boosted Tobit Models for Default Prediction.