WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Sales Prediction Using Machine Learning Techniques

Prajwal Amrutkar[1], Shubhangi Mahadik[2]

Bharti Vidyapeeth's Institute of Management and Technology

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, Objectives of sales prediction.In Section IV. The predictive analytics and methodology on sales price also discussed.

II. LITERATURE REVIEW

In this section, we will briefly review the previous studies on sales prediction and several classic prediction models. More than 200 kinds of prediction methods have been developed, which can be divided into two categories, subjective and objective methods. [2]
The subjective prediction method is based on the experience of experts who judge and estimate. It is strongly subjective and flexible. Examples are the Delphi method (Linstone & Turof, 1975), the brain storm method (Tremblay, Grosskopf, & Yang, 2010), the subjective probability method (Hogarth, 1975), and so on. These methods use the experience of experts or the integration of predicted results. In contrast, the objective prediction method uses raw data to build models based on mathematics and mathematical statistics methods[3]. It is reusable but not flexible. The objective prediction method includes mainly regression analysis and time series analysis. These methods use actual sales data, establishing a reusable model in order to predict future sales. Regression analysis methods include a simple regression model, a multivariate regression model, etc.

(Kleinbaum, Kupper, Nizam, et al., 2013). The time series analysis forecast model includes the moving average model, the exponential smoothing model, the seasonal trends model, the autoregressive-moving-average model, the generalized autoregressive conditional heteroscedastic model, etc. (Box, Jenkins, & Reinsel, 2013)

Most conventional sales prediction methods introduce either factors or time series to determine the forecast. McElroy and Burmeister (1988) applied Arbitrage Pricing Theory into a multivariate regression model. Lee and Fambro (1999) used the autoregressiveintegrated-moving-average model to do traffic volume forecasting. In 2003, Huang and Shih (2003) forecast shortterm loans using ARMA. Tay and Cao (2001) studied time series forecasting.

However, the relationship between influence factors or past time series data and sales prediction results is quite complicated. Therefore the predictions obtained from the aforementioned methods are often not satisfactory. As a consequence, many new intelligent model methods have recently been put to use in the area of forecasting; these perform better in terms of control and recognition. Some of the most representative new models are, for example, artificial neural networks (ANN) and support vector machines (SVM), the hot spots of forecasting research in recent of years. Kuo and Xue (1998) put forward a decision support system for sales prediction using fuzzy neural networks. Hill, Marquez, and O'Connor (1994) reviewed the artificial neural network models for forecasting and decision making. Cao (2003) combined SVM with time series for sales prediction while Gao et al. (2014) recommend extreme learning machine for sales prediction. Finally, Yuan (2014) proposed an online user behavior-based data mining method to predict sales in ecommerce.

However, the above research focused mainly on improving the accuracy of sales prediction via optimizing a single model algorithm or analyzing the factors that influence sales[4]. For special cases, such as when the sales volume was zero, the single prediction model didn't perform well. In addition, most of the previous methods only predicted results for one object, for example, one kind of book's sales. In actual situations, the approach needs to cover a large scale of products. Thus, the traditional single model optimization method has significant limitations in sales prediction.

We built a trigger model system instead of depending on a single model algorithm. Based on data about factors that influence sales, "the system" triggers one of the prediction models discussed previously, leading to better prediction results than before. Also, our method can be used for a much larger scale of sales prediction. Therefore, we provide a new proposal for sales prediction research, which has been proven to be a significant improvement over past methods through our validation.

III. OBJECTIVES

- To predict the future sales.
- To determine the amount of product that will be required in future.
- To compare and evaluate the performance of prediction algorithms.
- To analyze the past sales data

IV. 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 efashion store, for the three consecutive years of sales data. To predict the sales of the efashion 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 stages involved in the data mining model include data understanding, preparation, modelling, evaluation and deployment

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

D. Forecasting and Trends

Trend forecasting is the process of using market research and consumer data to create predictions about customers' future buying habits and preferences. Trend forecasting provides product designers with insight that may help them design an item that their target audience likes and purchases.

E. PREDICTION

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 . 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).

In this study, we impleented three machine learning algorithms on the training dataset and the models are tested for the performance[15]. 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 [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

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 modeld, required for adding a new weak learner into the model, the weak learner is the decision tree [19].
predictive analytics in sales forecasting. The big data analysis and forecasting are measured as the vital fields in the modern business scenario.

V. 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 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

VI. 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

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 KMeans 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. Ar Xiv 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. 24182427). 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.