2018 International Conference on Computing, Power and Communication Technologies (GUCON)
Galgotias University, Greater Noida, UP, India. Sep 28-29, 2018

A Two-Level Statistical Model for Big Mart Sales Prediction

Kumari Punam[1]
M.Tech Student,
Computer Science and Engineering,
IIT (ISM), Dhanbad, Jharkhand
kumaripunam23@gmail.com

Rajendra Pamula[2]
Assistant Professor ,
Computer Science and Engineering,
IIT (ISM), Dhanbad, Jharkhand
praphulajn1@gmail.com

Praphula Kumar Jain[3]
Junior Research Fellow,
Computer Science and Engineering,
IIT (ISM), Dhanbad, Jharkhand
rajendrapamula@gmail.com

***Abstract*— Sales forecasting is an important aspect of different companies engaged in retailing, logistics, manufacturing, marketing and wholesaling. It allows companies to efficiently allocate resources, to estimate achievable sales revenue and to plan a better strategy for future growth of the company. In this paper, prediction of sales of a product from a particular outlet is performed via a two-level approach that produces better predictive performance compared to any of the popular single model predictive learning algorithms. The approach is performed on Big Mart Sales data of the year 2013. Data exploration, data transformation and feature engineering play a vital role in predicting accurate results. The result demonstrated that the two-level statistical approach performed better than a single model approach as the former provided more information that leads to better prediction.**

Index Terms*— Data mining, Machine learning, Continuous values prediction, Stacking, Sales Prediction.

I. INTRODUCTION

Sales is a lifeblood of each and every company and sales forecasting plays a vital role in conducting any business. Good forecasting helps to develop and improve business strategies by increasing the knowledge about the marketplace. A standard sales forecast looks deeply into the situations or the conditions that previously occurred and then, applies inference regarding customer acquisition, identifies inadequacy and strengths before setting a budget as well as marketing plans for the upcoming year. In other words, sales forecasting is sales prediction that is based on the available resources from the past. An in-depth knowledge of the past resources allows to prepare for the upcoming needs of the business and increases the likelihood to succeed irrespective of external circumstances. Businesses that treat sales forecasting as the primary step tend to perform better than those who don't.

In this paper, the ensemble [1] of data mining predictive techniques via stacking is considered a two-level statistical approach. It is named as two-level because stacking is performed on two layers in which bottom layer consists of one or more than one learning algorithms and top layer consists of one learning algorithm. Stacking is also known as Stacked Generalization. It basically involves the training of the learning algorithm present in the top layer to combine the predictions made by the algorithms present in the bottom layer. In the first step, all the learning algorithms are trained using the big mart dataset and in the second step, a combiner algorithm is trained using all the predictions made by the bottom layer algorithms to get a final prediction. Stacking performs better than any single model because a stacking involves more information for prediction [2]. Each algorithm at lower level operates differently and makes individual predictions. All aspects of the dataset are viewed using various learning algorithms and then the final prediction is made.

To analyze the performance of the models, Mean Absolute Error (MAE) has been used as an evaluation metric. It is a metric to measure the accuracy for continuous variables. It is an average of absolute errors in a set of predictions, without considering their direction. Absolute error is the difference between the predicted value and the actual value.

This paper is sequenced in the following pattern: Section 1 presents a brief introduction to sales forecasting, stacking and the metric used for evaluation. Section 2 describes the related work that has been performed. Section 3 comprises of the detailed analysis of the steps that the proposed system undergoes. Section 4 demonstrates the experimental results and describes how the experiment was conducted and finally, the paper comes to an end with a conclusion in Section 5.

II. RELATED WORK

Machine learning [3] is the study of computational and statistical methods that automate the process of knowledge acquisition from experiences. Many types of research have been conducted on sales prediction and some of them has been analyzed and discussed below:

In paper [4], different machine learning algorithms has been discussed that were applied to different sectors of the industry like retailing, logistic marketing etc. as per the requirement. Beneficial knowledge about machine learning techniques is mentioned in the paper. It draws a conclusion that Rule Induction (RI) is the widely used machine learning technique in data mining applications in business compared to other four paradigms [5] in machine learning. The work performed in [6] describes the sale prediction of a pharmaceutical distribution company. The paper addresses two issues, firstly it performs stock proration so that it does not undergo out of stock state and secondly it targets sales prediction that manages the level of stock of medicines the company must keep in order to avoid any customer dissatisfaction. In paper [7] the objective is to handle the fluctuation of sales of footwear that occurs over time. It focuses on predicting weekly retail sales using the neural network. It reduces the uncertainty that exists in the short term planning of sales. A comparative analysis of linear and non-linear models to predict sales in the retailing sector is proposed in paper [8]. In paper [9], sales forecasting in fashion market is performed. Consumer-oriented markets face uncertain demands, lack of historical data and short life cycles. These factors challenge the forecasting methods to produce accurate results and preferably hybrid forecasting

978-1-5386-4491-1/18/$31.00 ©2018 IEEE

models perform better. A two-stage dynamic approach to make predictions in fashion retail is proposed in paper [10]. Short term as well as long-term predictions combined and produced a final prediction.

III. THE PROPOSED SYSTEM

The flow diagram in fig. 1 shows the sequence of steps that the dataset of Big Mart sales goes through to build up the proposed model to produce accurate results. There are a total of seven steps and each step plays a crucial role to build up the proposed model i.e. a two-level statistical model. The initial five steps is a pre-processing phase before building up the single model as well as the stack model and after the model building is done the model is tested with unseen data that measures the accuracy of the model. Smaller the value of the mean absolute error, better the model.

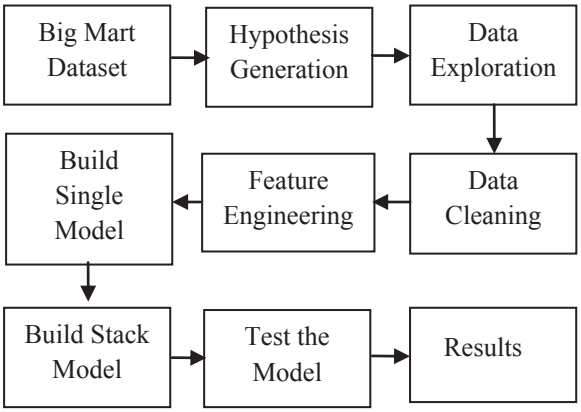

Fig. 1. Flow Diagram of the Proposed System

A. Big Mart Dataset Description

Big Mart is an International Retail Corporation. Big Mart Sales data of the year 2013 has been used as the dataset for the proposed work. Big Mart sales data has 12 attributes: Item Identifier, Item Fat, Item Visibility, Item Type, Outlet Type, Item MRP, Outlet Identifier, Item Weight, Outlet Size, Outlet Establishment Year, Outlet Location Type, and Item Outlet Sales. Item Outlet Sales is the response variable and rest of the attributes play the role of predictor variables. The dataset has 8523 observations. The dataset was partitioned into two parts, training data and test data in 80:20 ratio i.e. 6818 instances are training data and the rest 1705 instances are test data.

B. Hypothesis Generation

This is a very important step to analyze any problem. The first and foremost step is to have an understanding of the problem statement. The idea is to find out the factors of a product as well as of store (outlet) that creates an impact on the sales of a product. Many more attributes that are not present in the data can also be useful, it might not be sufficient but we can have a better understanding of the problem.

C. Data Exploration

After having a look at the dataset, certain information about the data was explored: Item weight and Outlet size have missing values. Item Visibility has a minimum value of zero that is practically not possible as if there is an object then it has to occupy some space. So, item visibility can never be zero. Outlet establishment year vary from 1985 to 2009. The values may not be appropriate in this form. So, we converted them to how old that particular outlet is 1559 unique products, as well as 10 unique outlets, are present in the dataset. Item type contains 16 unique values. Two types of Item Fat Content are there but some of them are misspelled as 'low fat', 'LF' instead of 'Low Fat and 'regular' instead of 'Regular'. From fig. 2 it was observed that the response variable i.e. Item Outlet Sales was positively skewed. So, log operation was performed on the response variable to remove skewness.

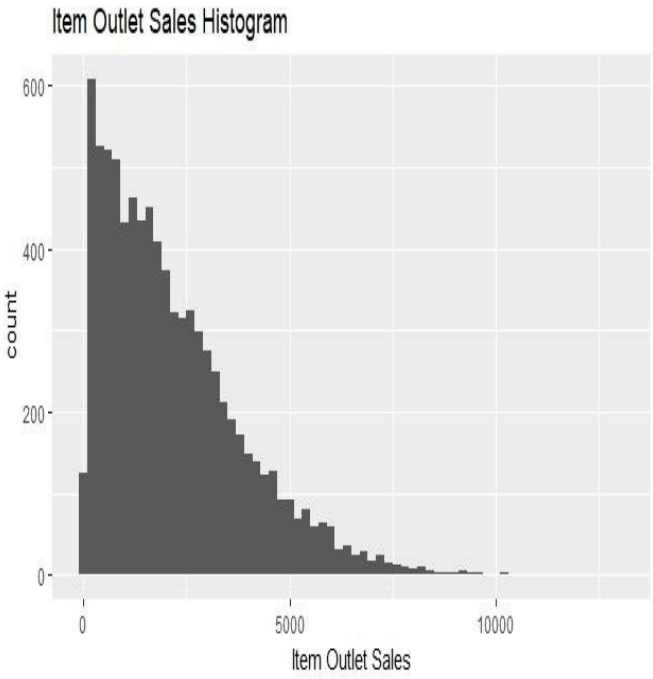

Fig. 2. Item Outlet Sales

D. Data Cleaning

In the data exploration phase, it was observed that Item Weight and Outlet Size had missing values. Missing values of Item Weight was filled by averaging the weight of that particular item and missing values of the Outlet Size was filled by the use of the mode of the outlet size for that particular type of outlet.

E. Feature Engineering

In the data exploration phase, some nuances in the dataset were discovered. So, in this phase, all of the nuances were resolved and made the data appropriate for modelling purpose. It was also observed that Item visibility had a zero value. So, these zero values were imputed with mean visibility of that particular product. Item fat content is resolved by changing all the miscoded ones to appropriate ones. It was seen that in some cases non-consumables possessed the fat content property which is practically not possible. So, a third category of Item fat content i.e. 'none' was created. In the Item Identifier attribute, it was observed that each of the unique ID started with either FD or DR or NC. So, a new column 'Item Type New' that had three categories Foods, Drinks and Non-consumables was created. One more new column 'Year' is added to the dataset that determines how much old that particular outlet is.

F. Model Building

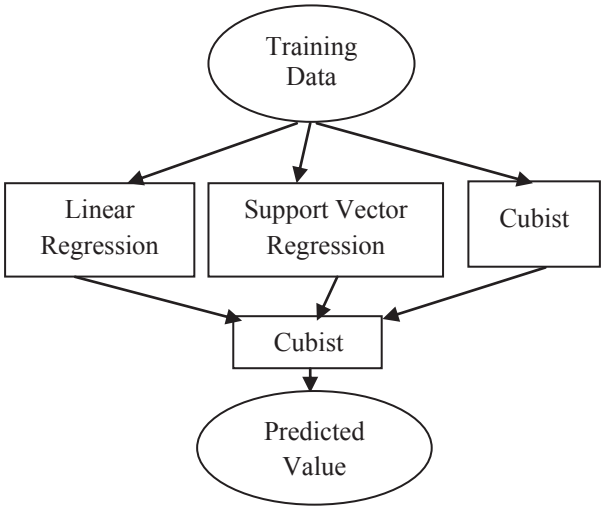

Fig. 3. A Two-level Statistical Model

After completing the phases of data exploration, data cleaning and feature engineering, the dataset is ready and set to build predictive models. Model building is a process of building a model that best explains the relationship between predictor variable and response variable. In this paper, model building takes place in two stages. At the first stage, the single model of popular predictive techniques like linear regression, regression tree [11], cubist [12], [13] and [14], support vector regression [15] and k- nearest neighbor [16] is built. Then in the second stage, the two-level statistical model was built. The two-level statistical model consisted of machine learning techniques such as linear regression, support vector regression and cubist. These machine learning algorithms are combined and used to make the final prediction. Stacking is a type of ensemble method that is generally used to combine the machine learning techniques to improve the accuracy of predictive models. It is basically a combination of different models that are treated as a single unit. Stacking may have more than two layers, it increases the complexity of the model but it may be useful to make accurate predictions.

From fig. 3, it is seen that, in the two-level statistical model, linear regression, support vector regression and cubist act as bottom layer models which take the original features of the dataset as inputs and then cubist act as the top layer model that takes the predictions of the bottom layer models as its input and makes the final prediction.

IV. EXPERIMENTAL RESULTS

To achieve the experimental results, each regression model employs a 10-fold cross-validation to appraise the predictive accuracy. In cross-validation, the big mart dataset is randomly partitioned into 10 subsets and each subset has roughly equal size. 9 subsets of the dataset form the training data and the left out subset are treated as the test data. The regression techniques form regression models using the training data and the test data is used for measuring the predictive accuracy. This process continues until every subset is treated as test data once. After analyzing the data via data visualization, it was seen that smallest location produced lowest sales. However, it was not same for the largest locations. From fig. 4, it was concluded that highest sales were not made from largest locations instead OUT027 produced highest sales whose size was medium and it was a supermarket of Type 3. So, to have an increment in the sales of the product from a particular outlet, Big Mart should switch more locations to Supermarket of Type 3. However, the two-level statistical model that single model should be good to predict future sales at its locations.

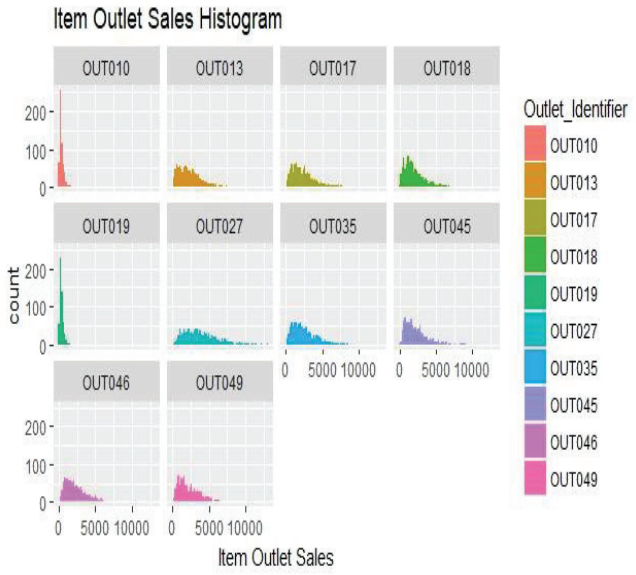

Fig. 4. Item Outlet Sales by Outlet Identifier

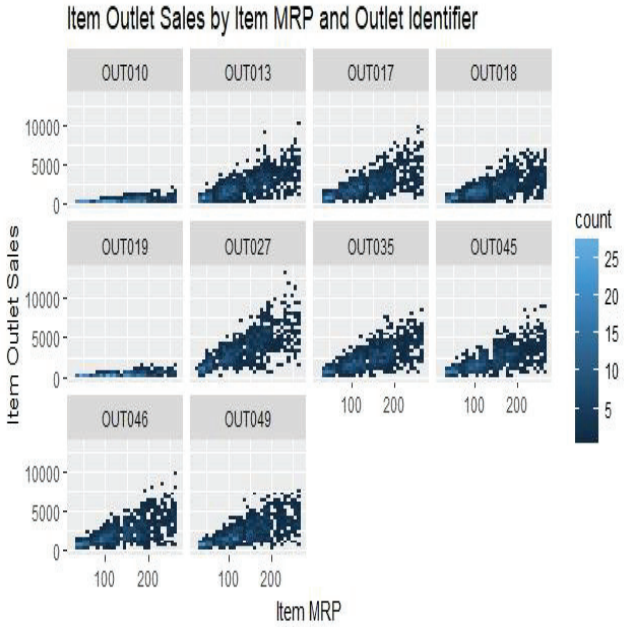

Fig. 5. Item Outlet Sales by Item MRP and Outlet Identifier

From fig. 5, it is concluded that as the price of the product increased, the sales of the product also increased.

From fig. 6, it is observed that Item Outlet Sales is highly correlated to Item MRP. These two variables share a linear relationship.

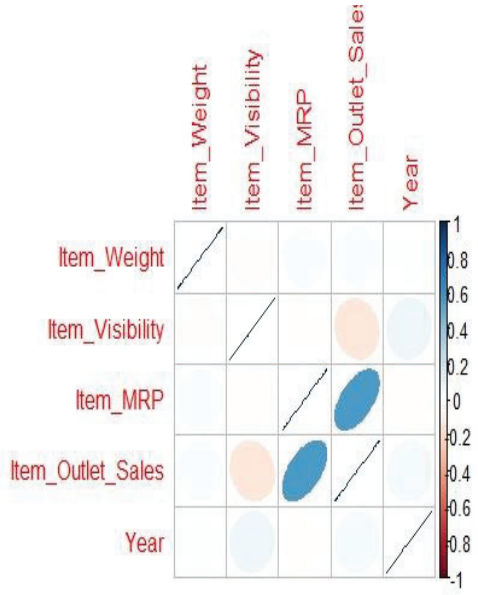

Fig. 6. Correlation Matrix

TABLE I. CROSS-VALIDATION MAE AND TEST MAE OF MACHINE LEARNING MODELS

Machine Learning Models	**Cross-Validation MAE**	**Test MAE**
k-NN	0.4573	0.4699
Linear Regression	0.4191	0.4142
Regression Tree	0.4184	0.413
Support Vector Regression	0.4144	0.4102
Cubist	0.4016	0.4000
Two-level Statistical Model	0.3975	0.3917

From Table I, it is concluded that the cross-validation MAE and the test MAE are within permissible limits and the two-level statistical model outperformed the rest of the single model predictive techniques.

V. CONCLUSION

Each and every company desires to know the demand of the customer in any season beforehand to avoid the shortage of products. As time passes by, the demand of the companies to be more accurate about the predictions will increase exponentially. So, huge research is going on in this sector to make accurate predictions of sales. Better predictions are directly proportional to the profit made by the company. Here in this paper, an effort has been made to predict sales of the product from a particular outlet accurately by using a two-level statistical model that reduces the mean absolute error value up to 39.17 %. The two-level statistical model outperformed the other single model predictive techniques and contributed better predictions to the big mart dataset.

REFERENCES

[1] Xu-Ying Liu, Jianxin Wu and Zhi-Hua Zhou, "Exploratory Undersampling for Class-Imbalance Learning", IEEE Transactions on Systems, Man and Cybernetics, Vol. 39(2), pp. 539-550, April, 2009.

[2] Deroski, Saso and Bernard Enko. "Is Combining Classifiers with Stacking Better Than Selecting the Best One?", Machine learning, Vol. 54(3), pp. 255-273, March, 2004

[3] Domingos Pedro, "A Few Useful Things to Know About Machine Learning", Communications of the ACM, Vol. 55, pp. 78-87, October, 2012.

[4] Bose, Indranil, and Radha K. Mahapatra, "Business Data Mining - A Machine Learning Perspective", Information & management, Vol. 39(3), pp. 211-225, February, 2001.

[5] Pat Langley and Herbert A. Simon, "Applications of Machine Learning and Rule Induction", Communications of the ACM, Vol. 38(11), pp. 54-64, November, 1995.

[6] Augusto Ribeiro, Isabel Seruca, and Natrcia Duro, "Improving Organizational Decision Support: Detection of Outliers and Sales Prediction for a Pharmaceutical Distribution Company", Procedia Computer Science, Vol. 121, pp. 282-290, December, 2017.

[7] Prasun Das and Subhasis Chaudhury, "Prediction of Retail Sales of Footwear using Feedforward and Recurrent Neural Networks", Neural Computing and Applications, Vol. 16(4), pp. 491-502, May, 2007.

[8] Ching-Wu Chu and Guoqiang Peter Zhang, "A Comparative Study of Linear and Nonlinear Models for Aggregate Retail Sales Forecasting", International Journal of Production Economics, Vol. 86(3), pp. 217-231, December, 2003.

[9] Beheshti-Kashi and Samaneh, "A Survey on Retail Sales Forecasting and Prediction in Fashion Markets", Systems Science & Control Engineering, Vol. 3, pp. 154-161, December, 2014.

[10] Yanrong Ni and Feiya Fan,"A Two-Stage Dynamic Sales Forecasting Model for the Fashion Retail", Expert Systems with Applications, Vol. 38(3), pp. 1529-1536, March, 2011.

[11] Wei-Yin Loh, "Classification and Regression Trees", Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, Vol. 1(1), pp. 14-23, January, 2011.

[12] J.R. Quinlan, "Learning with Continuous Classes", Proceedings of the 5th Australian Joint Conference on Artificial Intelligence, pp. 343-348, Tasmania, November, 1992.

[13] J.R. Quinlan, "Combining Instance Based and Model Based Learning", Proceedings of the Tenth International Conference on Machine Learning, pp. 236-243, San Mateo, June, 1993.

[14] Yong Wang and Ian H. Witten, "Inducing Model Trees for Continuous Classes", Proceedings of the Ninth European Conference on Machine Learning, pp. 128-137, Czech Republic, April, 1997.

[15] Alex J. Smola and Bernhard Schlkopf, "A Tutorial on Support Vector Regression", Statistics and computing, Vol. 14(3), pp. 199-222, 2004.

[16] Padraig Cunningham and Sarah Jane Delany, "k-Nearest Neighbour Classifiers", Multiple Classifier Systems, Vol. 34, pp. 1-17, March, 2007.