A Comparative Study of Big Mart Sales Prediction

Gopal Behera(✉) and Neeta Nain

Malaviya National Institute of Technology Jaipur, Jaipur, India
{2019rcp9002,nnain.cse}@mnit.ac.in

Abstract. Nowadays shopping malls and Big Marts keep the track of their sales data of each and every individual item for predicting future demand of the customer and update the inventory management as well. These data stores basically contain a large number of customer data and individual item attributes in a data warehouse. Further, anomalies and frequent patterns are detected by mining the data store from the data warehouse. The resultant data can be used for predicting future sales volume with the help of different machine learning techniques for the retailers like Big Mart. In this paper, we propose a predictive model using Xgboost technique for predicting the sales of a company like Big Mart and found that the model produces better performance as compared to existing models. A comparative analysis of the model with others in terms performance metrics is also explained in details.

Keywords: Machine learning · Sales forecasting · Random forest · Regression · Xgboost

1 Introduction

Day by day competition among different shopping malls as well as big marts is getting more serious and aggressive only due to the rapid growth of the global malls and on-line shopping. Every mall or mart is trying to provide personalized and short-time offers for attracting more customers depending upon the day, such that the volume of sales for each item can be predicted for inventory management of the organization, logistics and transport service, etc. Present machine learning algorithm are very sophisticated and provide techniques to predict or forecast the future demand of sales for an organization, which also helps in overcoming the cheap availability of computing and storage systems. In this paper, we are addressing the problem of big mart sales prediction or forecasting of an item on customer's future demand in different big mart stores across various locations and products based on the previous record. Different machine learning algorithms like linear regression analysis, random forest, etc are used for prediction or forecasting of sales volume. As good sales are the life of every organization so the forecasting of sales plays an important role in any shopping complex. Always a better prediction is helpful, to develop as well as to

N. Nain et al. (Eds.): CVIP 2019, CCIS 1147, pp. 421–432, 2020.
https://doi.org/10.1007/978-981-15-4015-8_37

enhance the strategies of business about the marketplace which is also helpful to improve the knowledge of marketplace. A standard sales prediction study can help in deeply analyzing the situations or the conditions previously occurred and then, the inference can be applied about customer acquisition, funds inadequacy and strengths before setting a budget and marketing plans for the upcoming year. In other words, sales prediction is based on the available resources from the past. In depth knowledge of past is required for enhancing and improving the likelihood of marketplace irrespective of any circumstances especially the external circumstance, which allows to prepare the upcoming needs for the business. Extensive research is going on in retailers domain for forecasting the future sales demand. The basic and foremost technique used in predicting sale is the statistical methods, which is also known as the traditional method, but these methods take much more time for predicting a sales also these methods could not handle non linear data so to over these problems in traditional methods machine learning techniques are deployed. Machine learning techniques can not only handle non-linear data but also huge data-set efficiently. To measure the performance of the models, Root Mean Square Error (RMSE) [15] and Mean Absolute Error (MAE) [4] are used as an evaluation metric as mentioned in the Eqs. 1 and 2 respectively. Here Both metrics are used as the parameter for accuracy measure of a continuous variable.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |\mathrm{x}_{\mathrm{predict}} - \mathrm{x}_{\mathrm{actual}}| \tag{1}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (|\mathrm{x}_{\mathrm{predict}} - \mathrm{x}_{\mathrm{actual}}|)^2} \tag{2}$$

where n: total number of error and $|\mathrm{x}_{\mathrm{predict}} - \mathrm{x}_{\mathrm{actual}}|$: Absolute error. The remaining part of this article is arranged as following: Sect. 1 briefly describes introduction of sales prediction of Big Mart and also elaborate about the evaluation metric used in the model. Previous related work has been pointed in Sect. 2. The detailed description and analysis of proposed model is given in Sect. 3. Where as implementations and results are demonstrated in Sect. 4 and the paper concludes with a conclusion in the last section.

2 Related Work

Sales forecasting as well as analysis of sale forecasting has been conducted by many authors as summarized: The statistical and computational methods are studied in [2] also this paper elaborates the automated process of knowledge acquisition. Machine learning [6] is the process where a machine will learn from data in the form of statistically or computationally method and process knowledge acquisition from experiences. Various machine learning (ML) techniques with their applications in different sectors has been presented in [2]. Langley and Simon [7] pointed out most widely used data mining technique in the field

of business is the Rule Induction (RI) technique as compared to other data mining techniques. Where as sale prediction of a pharmaceutical distribution company has been described in [10,12]. Also this paper focuses on two issues: (i) stock state should not undergo out of stock, and (ii) it avoids the customer dissatisfaction by predicting the sales that manages the stock level of medicines. Handling of footwear sale fluctuation in a period of time has been addressed in [5]. Also this paper focuses on using neural network for predicting of weekly retail sales, which decrease the uncertainty present in the short term planning of sales. Linear and non-linear [3] a comparative analysis model for sales forecasting is proposed for the retailing sector. Beheshti-Kashi and Samaneh [1] performed sales prediction in fashion market. A two level statistical method [11] is elaborated for forecasting the big mart sales prediction. Xia and Wong [17] proposed the differences between classical methods (based on mathematical and statistical models) and modern heuristic methods and also named exponential smoothing, regression, auto regressive integrated moving average (ARIMA), generalized auto regressive conditionally heteroskedastic (GARCH) methods. Most of these models are linear and are not able to deal with the asymmetric behavior in most real-world sales data [9]. Some of the challenging factors like lack of historical data, consumer-oriented markets face uncertain demands, and short life cycles of prediction methods results in inaccurate forecast.

3 Proposed System

For building a model to predict accurate results the dataset of Big Mart sales undergoes several sequence of steps as mentioned in Fig. 1 and in this work we propose a model using Xgboost technique. Every step plays a vital role for building the proposed model. In our model we have used 2013 Big mart dataset [13]. After preprocessing and filling missing values, we used ensemble classifier using Decision trees, Linear regression, Ridge regression, Random forest and Xgboost. Both MAE and RSME are used as accuracy metrics for predicting the sales in Big Mart. From the accuracy metrics it was found that the model will predict best using minimum MAE and RSME. The details of the proposed method is explained in the following section.

3.1 Dataset Description of Big Mart

In our work we have used 2013 Sales data of Big Mart as the dataset. Where the dataset consists of 12 attributes like Item_ Fat, Item_Type, Item_MRP, Outlet_Type, Item_Visibility, Item_Weight, Outlet_Identifier, Outlet_Size, Outlet Establishment Year, Outlet_Location_Type, Item_Identifier and Item_Outlet_Sales. Out of these attributes response variable is the Item_Outlet_Sales attribute and remaining attributes are used as the predictor variables. The data-set consists of 8523 products across different cities and locations. The data-set is also based on hypotheses of store level and product level. Where store level involves attributes like: city, population density, store

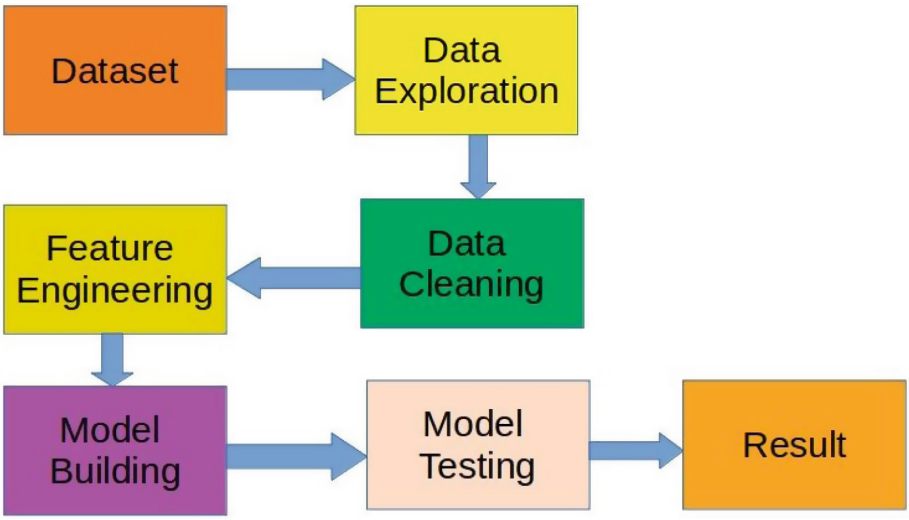

Fig. 1. Working procedure of proposed model.

capacity, location, etc and the product level hypotheses involves attributes like: brand, advertisement, promotional offer, etc. After considering all, a dataset is formed and finally the data-set was divided into two parts, training set and test set in the ratio 80:20.

3.2 Data Exploration

In this phase useful information about the data has been extracted from the dataset. That is trying to identify the information from hypotheses vs available data. Which shows that the attributes Outlet size and Item weight face the problem of missing values, also the minimum value of Item Visibility is zero which is not actually practically possible. Establishment year of Outlet varies from 1985 to 2009. These values may not be appropriate in this form. So, we need to convert them into how old a particular outlet is. There are 1559 unique products, as well as 10 unique outlets, present in the dataset. The attribute Item _ type contains 16 unique values. Where as two types of Item_Fat_ Content are there but some of them are misspelled as regular instead of 'Regular' and low fat', 'LF' instead of 'Low Fat'. From Fig. 2. It was found that the response variable i.e. Item_Outlet_Sales was positively skewed. So, to remove the skewness of response variable a log operation was performed on Item_Outlet_Sales.

3.3 Data Cleaning

It was observed from the previous section that the attributes Outlet Size and Item Weight has missing values. In our work in case of Outlet Size missing value we replace it by the mode of that attribute and for the Item Weight missing values we replace by mean of that particular attribute. The missing

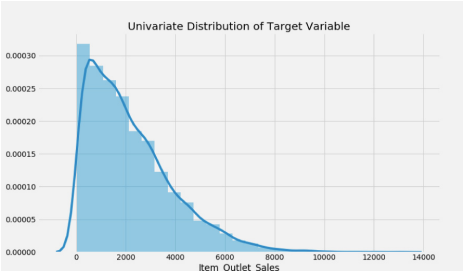

Fig. 2. Univariate distribution of target variable Item outlet sales. The Target variable is positively skewed towards the higher sales.

attributes are numerical where the replacement by mean and mode diminishes the correlation among imputed attributes. For our model we are assuming that there is no relationship between the measured attribute and imputed attribute.

3.4 Feature Engineering

Some nuances were observed in the data-set during data exploration phase. So this phase is used in resolving all nuances found from the dataset and make them ready for building the appropriate model. During this phase it was noticed that the Item_visibility attribute had a zero value, practically which has no sense. So the mean value item visibility of that product will be used for zero values attribute. This makes all products likely to sell. All categorical attributes discrepancies are resolved by modifying all categorical attributes into appropriate ones. In some cases, it was noticed that non-consumables and fat content property are not specified. To avoid this we create a third category of Item fat content i.e. "none". In the Item_Identifier attribute, it was found that the unique ID starts with either DR or FD or NC. So, we create a new attribute 'Item_Type_New' with three categories like Foods, Drinks and Non-consumables. Finally, for determining how old a particular outlet is, we add an additional attribute 'Year' to the dataset.

3.5 Model Building

After completing the previous phases, the dataset is now ready to build proposed model. Once the model is build it is used as predictive model to forecast sales of Big Mart. In our work, we propose a model using Xgboost algorithm and compare it with other machine learning techniques like Linear regression, Ridge regression [14], Decision tree [8,16] etc.

Decision Tree: A decision tree classification is used in binary classification problem and it uses entropy [8] and information gain [16] as metric and is defined in Eqs. 3 and 4 respectively for classifying an attribute which picks the highest information gain attribute to split the data set.

$$H(S) = -\sum_{c \in C} \mathrm{p(c)} \log \mathrm{p(c)} \tag{3}$$

where H(S): Entropy, C: Class Label, P: Probability of class c.

$$\text{Infromation Gain}(S, A) = H(S) - \sum_{t \in T} \mathrm{p(t)H(t)} \tag{4}$$

where S: Set of attribute or dataset, $H(S)$: Entropy of set S, T: Subset created from splitting of S by attribute A. $p(t)$: Proportion of the number of elements in t to number of element in the set S. $H(t)$: Entropy of subset t. The decision tree algorithm is depicted in Algorithm 1.

Require: Set of features d and set of training instances D
1: **if** *all the instances in D have the same target label C* **then**
  2: Return a decision tree consisting of leaf node with label level C
end
else if *d is empty* **then**
  4: Return a decision tree of leaf node with label of the majority target level in D
end
5: **else if** *D is empty* **then**
  6: Return a decision tree of leaf node with label of the majority target level of the immediate parent node
end
7: **else**
  8: $d[best] \leftarrow$ arg max IG(d, D) where $d \in D$
  9: make a new node, Node$_{d[best]}$
  10: partition D using $d[best]$
  11: remove $d[best]$ from d
  12: **for** *each partition* D_i *of* D **do**
    13: grow a branch from Node$_{d[best]}$ to the decision tree created by rerunning $ID3$ with D=D$_i$
  end
end

Algorithm 1: ID3 algorithm

Linear Regression: A model which create a linear relationship between the dependent variable and one or more independent variable, mathematically linear regression is defined in Eq. 5

$$y = \sum w^T x \tag{5}$$

where y is dependent variable and x are independent variables or attributes. In linear regression we find the value of optimal hyperplane w which corresponds to the best fitting line (trend) with minimum error. The loss function for linear regression is estimated in terms of RMSE and MAE as mentioned in the Eqs. 1 and 2.

Ridge Regression: The cost function for ridge regression is defined in Eq. 6.

$$min\left(|(Y - X(\theta)|)^2 + \lambda \|\theta\|^2\right) \tag{6}$$

here λ known as the penalty term as denoted by α parameter in the ridge function. So the penalty term is controlled by changing the values of α, higher the values of α bigger is the penalty. Figure 3 shows Linear Regression, Ridge Regression, Decision Tree and proposed model i.e. Xgboost.

Xgboost (Extreme Gradient Boosting) is a modified version of Gradient Boosting Machines (GBM) which improves the performance upon the GBM framework by optimizing the system using a differentiable loss function as defined in Eq. 7.

$$\sum_{i=1}^{n} l(\mathrm{y}_i, \hat{y}_i) + \sum_{k} k\Omega(f_k), f_k \in F \tag{7}$$

where $\hat{y_i}$: is the predicted value, y_i : is the actual value and F is the set of function containing the tree, $l(y_i, \hat{y_i})$ is the loss function.

This enhances the GBM algorithm so that it can work with any differentiable loss function. The GBM algorithm is illustrated in Algorithm 2.

Step 1: Initialize model with a constant value:

$$F_0 = arg\, min \sum_{i=0}^{n} L(y_i, \gamma)$$

Step 2: **for** *m= 1 to M :* **do**

a. Compute pseudo residuals:

$$r_{im} = -\left[\frac{\partial L(y_i F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)}$$

for all $i = 1, 2...n$

b. Fit a Base learner $h_m(x)$ to pseudo residuals that is train the learner using training set.

c. Compute γ_m

$$\gamma_m = arg\min_{\gamma} \sum_{i=0}^{n} (L(y_i, F_{m-1}(x_i) + \gamma h(x_i)))$$

d. Update the model:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x)$$

end

Step 3: Output F_M

Algorithm 2: Gradient boosting machine(GBM) algorithm

The Xgboost has following exclusive features:

1. Sparse Aware - that is the missing data values are automatic handled.
2. Supports parallelism of tree construction.
3. Continued training - so that the fitted model can further boost with new data.

All models received features as input, which are then segregated into training and test set. The test dataset is used for sales prediction.

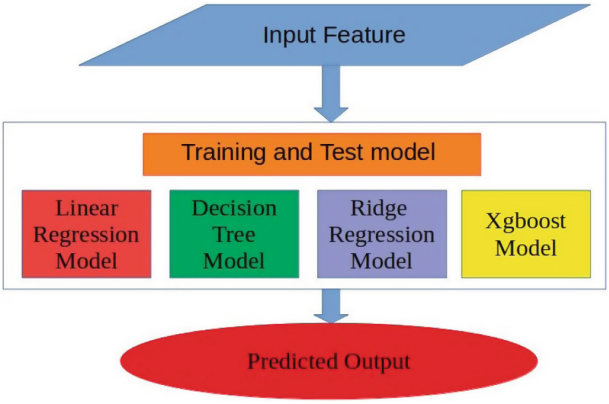

Fig. 3. Framework of proposed model. Model received the input features and split it into training and test set. The trained model is used to predict the future sales.

4 Implementation and Results

In our work we set cross-validation as 20 fold cross-validation to test accuracy of different models. Where in the cross-validation stage the dataset is divided randomly into 20 subsets with roughly equal sizes. Out of the 20 subsets, 19 subsets are used as training data and the remaining subset forms the test data also called leave-one-out cross validation. Every models is first trained by using the training data and then used to predict accuracy by using test data and this continues until each subset is tested once.

From data visualization, it was observed that lowest sales were produced in smallest locations. However, in some cases it was found that medium size location produced highest sales though it was type-3 (there are three type of super market e.g. super market type-1, type-2 and type-3) super market instead of largest size location as shown in Fig. 4. To increase the product sales of Big mart in a particular outlet, more locations should be switched to Type 3 Supermarkets.

However, the proposed model gives better predictions among other models for future sales at all locations. For example, how item MRP is correlated with outlet

sales is shown in Fig. 5. Also Fig. 5 shows that Item_Outlet_Sales is strongly correlated with Item_MRP, where the correlation is defined in Eq. 8.

$$P_{Corr} = \frac{n\sum(xy) - (\sum x)(\sum y)}{\sqrt{n[\sum x^2) - (\sum x)^2]}\sqrt{n[\sum y^2 - (\sum y)^2]}} \tag{8}$$

From Fig. 8 it is also observed that target attribute Item_Outlet_Sales is affected by sales of the Item_Type. Similarly, from Fig. 6 it is also observed that highest sales is made by OUT027 which is actually a medium size outlet in the super market type-3. Figure 7 describes that the less visible products are sold more compared to the higher visibility products which is not possible practically. Thus, we should reject the one of the product level hypothesis that is the visibility does not effect the sales.

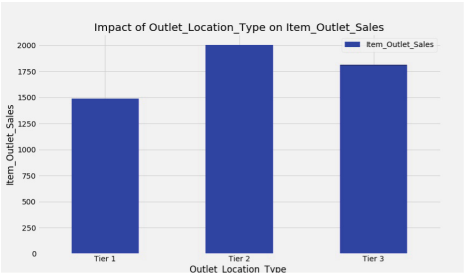

Fig. 4. Impact of outlet location type on target variable item outlet sale. Displayed the sales volume of different outlet locations.

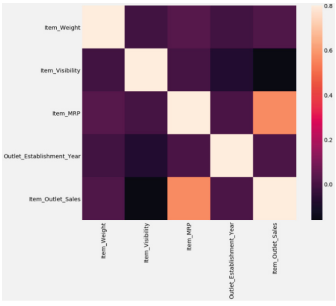

Fig. 5. Correlation among features of a dataset. Brown squares are highly correlated whereas black square represents bad correlation among attributes. (Color figure online)

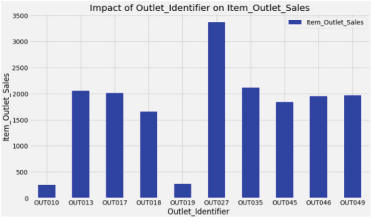

Fig. 6. Impact of outlet identifier on target variable item outlet sale.

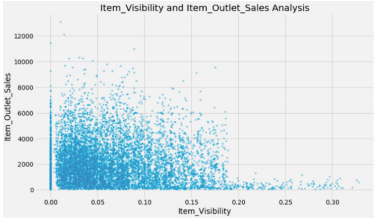

Fig. 7. Impact of item visibility on target variable item outlet sale. Less visible items are sold more compared to more visibility items as outlet contains daily used items which contradicts the null hypothesis.

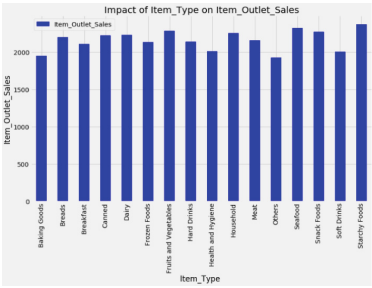

Fig. 8. Impact of item type on target variable item outlet sale.

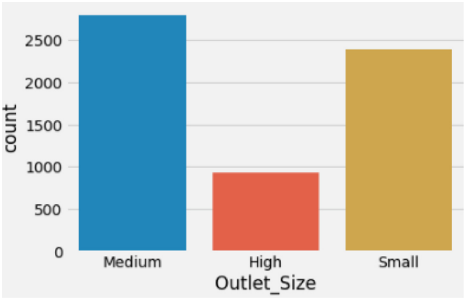

Fig. 9. Distribution of outlet size. The number of outlet size are available in the dataset.

From Fig. 9 it is observed that less number of high outlet size stores exist in comparison to the medium and small outlet size in terms of count. The cross-validation score along with MAE and RMSE of the proposed model and existing models is shown in Tables 1 and 2 respectively. Similarly the root mean squared error for existing model and proposed model is presented in Table 2. From the results we observe that and found that the proposed model is significantly improved over the other model.

Table 1. Comparison of cross validation score of different model

Model	Cross validation score (Mean)	Cross validation score (Std)
Linear regression	1129	43.24
Decision tree	1091	45.42
Ridge regression	1097	43.41

Table 2. Comparison of MAE and RMSE of proposed model with other model

Model	MAE	RMSE
Linear regression	836.1	1127
Decision tree	751.6	1068
Ridge regression	836	1129
Xgboost	**749.03**	**1066**

5 Conclusions

In present era of digitally connected world every shopping mall desires to know the customer demands beforehand to avoid the shortfall of sale items in all seasons. Day to day the companies or the malls are predicting more accurately the

demand of product sales or user demands. Extensive research in this area at enterprise level is happening for accurate sales prediction. As the profit made by a company is directly proportional to the accurate predictions of sales, the Big marts are desiring more accurate prediction algorithm so that the company will not suffer any losses. In this research work, we have designed a predictive model by modifying Gradient boosting machines as Xgboost technique and experimented it on the 2013 Big Mart dataset for predicting sales of the product from a particular outlet. Experiments support that our technique produce more accurate prediction compared to than other available techniques like decision trees, ridge regression etc. Finally a comparison of different models is summarized in Table 2. From Table 2 it is also concluded that our model with lowest MAE and RMSE performs better compared to existing models.

References

1. Beheshti-Kashi, S., Karimi, H.R., Thoben, K.D., Lütjen, M., Teucke, M.: A survey on retail sales forecasting and prediction in fashion markets. Syst. Sci. Control Eng. **3**(1), 154–161 (2015)
2. Bose, I., Mahapatra, R.K.: Business data mining-a machine learning perspective. Inf. Manage. **39**(3), 211–225 (2001)
3. Chu, C.W., Zhang, G.P.: A comparative study of linear and nonlinear models for aggregate retail sales forecasting. Int. J. Prod. Econ. **86**(3), 217–231 (2003)
4. Claypool, M., Gokhale, A., Miranda, T., Murnikov, P., Netes, D., Sartin, M.: Combing content-based and collaborative filters in an online newspaper (1999)
5. Das, P., Chaudhury, S.: Prediction of retail sales of footwear using feedforward and recurrent neural networks. Neural Comput. Appl. **16**(4–5), 491–502 (2007). https://doi.org/10.1007/s00521-006-0077-3
6. Domingos, P.M.: A few useful things to know about machine learning. Commun. ACM **55**(10), 78–87 (2012)
7. Langley, P., Simon, H.A.: Applications of machine learning and rule induction. Commun. ACM **38**(11), 54–64 (1995)
8. Loh, W.Y.: Classification and regression trees. Wiley Interdisc. Rev. Data Min. Knowl. Disc. **1**(1), 14–23 (2011)
9. Makridakis, S., Wheelwright, S.C., Hyndman, R.J.: Forecasting Methods and Applications. Wiley, New York (2008)
10. Ni, Y., Fan, F.: A two-stage dynamic sales forecasting model for the fashion retail. Expert Syst. Appl. **38**(3), 1529–1536 (2011)
11. Punam, K., Pamula, R., Jain, P.K.: A two-level statistical model for big mart sales prediction. In: International Conference on Computing, Power and Communication Technologies (GUCON), pp. 617–620. IEEE (2018)
12. Ribeiro, A., Seruca, I., Durão, N.: Improving organizational decision support: detection of outliers and sales prediction for a pharmaceutical distribution company. Procedia Comput. Sci. **121**, 282–290 (2017)
13. Shrivas, T.: Big mart dataset@ONLINE, June 2013. https://datahack.analyticsvidhya.com/contest/practice-problem-big-mart-sales-iii/
14. Smola, A.J., Schölkopf, B.: A tutorial on support vector regression. Stat. Comput. **14**(3), 199–222 (2004). https://doi.org/10.1023/B:STCO.0000035301.49549.88
15. Smyth, B., Cotter, P.: Personalized electronic program guides for digital TV. AI Mag. **22**(2), 89 (2001)

16. Wang, Y., Witten, I.H.: Induction of model trees for predicting continuous classes (1996)
17. Xia, M., Wong, W.K.: A seasonal discrete grey forecasting model for fashion retailing. Knowl.-Based Syst. **57**, 119–126 (2014)