Smartphone Price Prediction in Retail Industry Using Machine Learning Techniques

K. T. Chandrashekhara, M. Thungamani, C. N. Gireesh Babu and T. N. Manjunath

Abstract The goal of any organization is to make their product to get succeed and compete with other products in the market where pricing of their products plays a vital role. To sell any product in market, the most important aspect is to determine the price. There are many traditional and new methods for estimating before pricing their products, and a method is chosen which gives more appropriate result. In this study, support vector regression analysis is used as a machine learning technique in order to predict the market price of smartphones based on their features. Many variants of features are utilized for data preprocessing or input technique for SVR model. If required factors are derived and used accordingly, it can provide a good prediction result. Different features of the smartphone are considered in this experiment in order to get more reliable outputs. Support vector regression gives more promising predictions for making better decisions in price prediction of smartphones compared to other models.

Keywords Support vector machine (SVM) · Machine learning · Decision making · Price prediction

1 Introduction

"In current market situations, customers do not purchase products only based on price alone." In the event that a retailer is comfortable in choosing the cost for his items, it is conceivable in current markets with the present pricing techniques [1]. The competition within the smartphone industry sector is growing day by day to an entirely new level and with modern advanced technology growing every minute with modern smartphone companies like Vivo, Mi, Samsung, Apple, and Vibe. The

K. T. Chandrashekhara (✉) · C. N. Gireesh Babu · T. N. Manjunath
Department of ISE, BMSIT&M, Bangalore, India
e-mail: chandru.kt@gmail.com

M. Thungamani
University of Horticultural Sciences, Bagalkot, India

V. Sridhar et al. (eds.), *Emerging Research in Electronics, Computer Science and Technology*, Lecture Notes in Electrical Engineering 545,
https://doi.org/10.1007/978-981-13-5802-9_34

New innovations discovers the price factor estimations of different products of the individual firms. The companies come with new ideas and technology innovations to survive in the market and to compete with other companies. This is where the data analytics and its approaches come into picture [2]. Smartphone businesses are considered to be one of the evergreen fields, because an item with new trending and innovation can make a huge amount of profit, not only for smart-phones, estimating analytics be used for each item and products of the companies. Pricing analytics strategies require the past history and data to work on and predict the results; there are lots of factors which need to be taken care of predictive factors such as manufacturing expenses which include the cost of each part used, labor charges, marketing charges, importing or exporting charges, advertising, and promotions of the product [3]. It is also needed to maintain the quality and standards of the products to survive in the market. The first thing buyer looks is at the cost of the product which makes price an important factor. Even with average quality and less specifications, a good price can overcome the disadvantages and make the product successful one. Other factors that are required for the process predicting the price are market condition, brand, the quality of the product, and the actual demand for the product [4].

2 Existing System

Huge amounts of data have been generated, from the last two decades from various business domains such as stock markets, sensor devices, temperature sensors, social media, retail industry, and health care. However, these data are not useful without proper analytic power. Numerous big data solutions have enabled people to obtain meaningful data insight into a large amount of data generated by various business domains. However, these solutions are still in their infancy, and the domain lacks a comprehensive survey. Big data analytics in IOT requires processing a large amount of data on the fly and storing the results in various storage technologies. The relationship between big data and Internet of things can be managed in three phases. In the first phase, different types of devices generate huge amounts of data with different data forms, and this data can be stored in the low-cost cloud storages. In the second phase, the generated data are known as “big data,” which is based on volume, velocity, and variety stored in big data file system. In the last phase, we apply various analytical tools such as MapReduce, Storm, Spark to analyze the stored big data, datasets.

With the rapid increase in the amount of data generation, the process of selecting machine learning tools for big data analytics is very difficult. The existing tools have advantages and disadvantages, and many have overlapping uses. The world’s data are growing at faster rate, and traditional tools for machine learning are becoming inefficient as we move toward faster distributed and real-time processing. Here, we need to aid the researcher or professional who understands machine learning but is inexperienced with big data. In order to evaluate machine learning tools, one should have a thorough understanding of what to look for. To that extent, here we provide a list of criteria for making selections along with an analysis of the advantages and drawbacks of each. We can do this by starting from the beginning and looking at what exactly the term "big data analytics" means.

The multilayer feed forward neural network is used in the existing system, which is used to classify the input data based on the contribution to the value of the product. The multilayer feed forward network consists of one input layer, one hidden layer, and one output layer [5]. The output layer consists of three neurons, which represent the quality of the product under three categories that are brand name, outlook, and specification. These three output values will use to predict the unknown price values. Once the input data is converted to the required format, the neuron structure is outlined by setting the basic parameters like activation function, threshold value, and learning rate. The learning rate is actually calculated at runtime.

3 Proposed System

Here, we are using three machine learning algorithms for price prediction, namely

1. Support vector regression (SVR)
2. Backpropagation neural network (BNN)
3. Linear regression (LR).

These algorithms are applied to the dataset for price prediction. Based on the results, we have done a comparative study of which algorithm is more efficient for this case.

3.1 Dataset

Our dataset is consisting of the historical data of the smartphones that are accessible over the last few years disregarding of location of producing and cost of producing, so that we are able to effectively correlate with effective output. The retail market value of smartphones is the main objective of our study. Dataset not solely records the retail cost of smartphones, but records each alternative side like specifications, name, features. Dataset not just records the retail cost of phones, but records each other perspective like name, features. Our analysis requires totally different aspects of analytics to get the accuracy in our cost prediction. There are totally 262 records in the dataset. The recordings date from March 2010 to January 2018. These datasets are obtained through e-commerce sites. Eighty percentage of the data (210) are used as training data, and remaining 20% (52) are used as test data (Fig. 1).

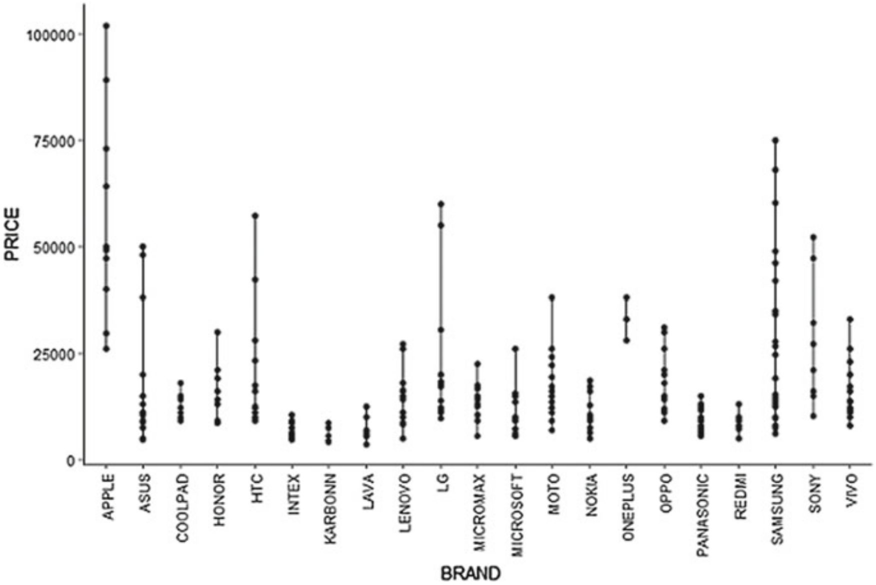

Fig. 1 Graphical representation of input data

3.2 Feature Selection

The price of smartphones depends upon its features. Here, we have considered features like brand, OS, RAM, memory, weight, dimension, display, technologies, primary camera, secondary camera, processor, color, and battery power to predict the price. The summary of features is as follows. There are about 21 different brands of phones with 3 different OS (Android, Windows, and iOS). RAM size varies from 0.5 to 8 GB. Memory ranges from 4 to 256 GB. The average weight of the phones is 172 g. Dimension is measured in cm^3. The mean display size is 5.2 inches, and the average battery power is 2900 mAh. Technology is nothing but 2G, 3G, or 4G. Processors used can be dual, quad, hexa, octa, or deca cores. There are mainly 14 different colors of phones. Price varies from 3500 to 100,000 Rs.

3.3 Methodology

We have collected the data of smartphones from e-commerce Web sites. We are storing the dataset in CSV format. After reading data, we are doing data preprocessing for required fields to improve the resulting efficiency. Mainly, we have eliminated duplicated entries, dealt with NAs and null values. For example, we have converted

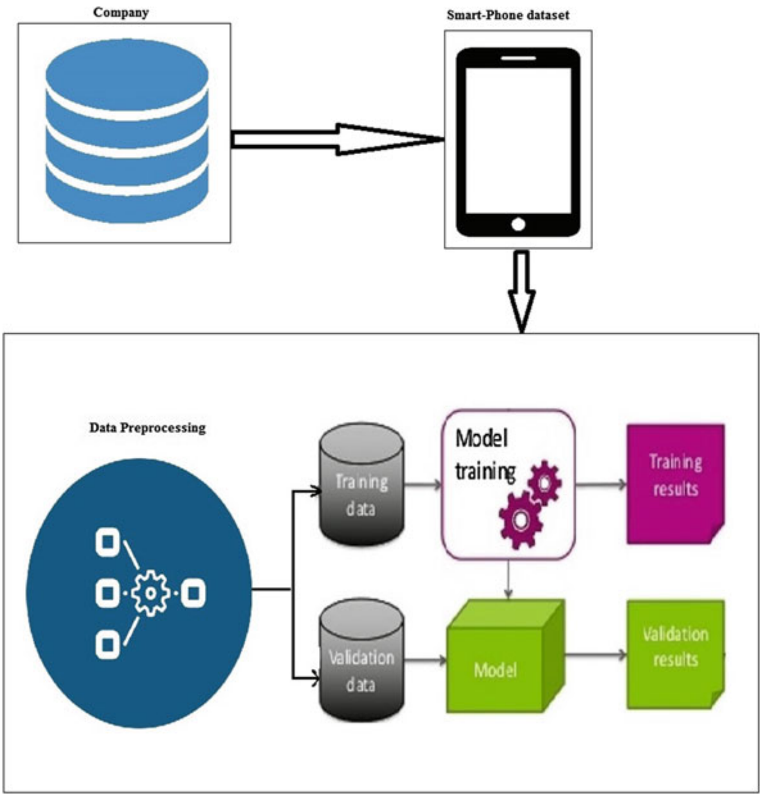

Fig. 2 System architecture

all letters to upper case to maintain uniformity in the dataset. We also converted factors into numeric values as many of our models deal with numerical values.

Preprocessed data are divided into training data and testing data. The training data are used to train the models, and the testing data are used for validating the results. Given the features of a smartphone, our trained model can predict the price for it (Fig. 2).

3.4 Models

Support vector machine—SVM is a supervised machine learning algorithm that can be used for classification and for regression purposes. Since our response variable is a real number, we are using support vector regression [6]. First of all, because the output is a real number, it becomes very difficult to predict the information at hand, which has infinite possibilities. Since some of the input features have a nonlinear relation with the output variable, kernel function is used to transform the data into a higher-dimensional feature space to make it possible to perform the linear separation. Here, the radial basis kernel function (K) is used as it produced a good result compared to other kernel functions.

Algorithm:

```
Library e1071
Model = svm (PRICE~., TrainData, kernel=″radial″,
type=″eps-regression)
Result = predict (Model, TestData)
```

Here to run SVM algorithm, we are using e1071 package. Function svm() of this package is used to train the model. Function predict() is used to predict price for testing dataset.

$$y = \sum_{i=1}^{n} (\alpha i - \alpha i^{*}) \cdot K(xi, x) + b \tag{1}$$

$$K\left(x_i, x_j\right) = \exp\left(-\frac{|xi - xj|^2}{2\sigma^2}\right) \tag{2}$$

$$\text{The idea is to minimize } \frac{1}{2}|\omega|^2 \tag{3}$$

$$\begin{aligned} \text{Subjected to constraints } yi - \omega x_i - b &\le \varepsilon \\ \omega x_i + b - y_i &\le \varepsilon \end{aligned} \tag{4}$$

Multiple linear regression—This is another machine learning algorithm used to train the model. It finds the linear relation between input variables and output variable by fitting a linear equation.

Algorithm:

```
Model = lm (PRICE~., TrainData)
Result = predict (Model, TestData)
```

Here, lm() is the function used for training and predict() is the function for testing the data.

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon_i \tag{5}$$

where

y_i is actual value of each y (response variable),
x_i $x_1, x_2, \ldots x_n$ are independent variables (predictors),
β_0 is the intercept,
$\beta_1, \beta_2, \ldots, \beta_n$ are regression coefficients,
ϵ_i is the residual error.

$$\mathrm{\hat{y}_i = b_0 + b_1x_1 + b_2x_2 + \cdots + b_nx_n} \tag{6}$$

where $\hat{y}_i$ is the predicted value (fitted) of each y and b_0, b_1, …, b_n are the estimates of $\beta_0, \beta_1, \beta\rho$

$$\mathrm{\epsilon_i = y_i - \hat{y}_I} \tag{7}$$

Backpropagation neural network—We have also implemented the existing backpropagation neural network model for comparison of results. Here, the network consists of an input layer with 11 neurons, one hidden layer, and an output layer with one neuron representing output variable price. Min–max normalization has been carried out to get more efficient results.

Algorithm:

```
Data = scale (Dataset)
Library neuralnet
Model = neuralnet (PRICE~., Traindata, hidden = 1)
Result = compute (Model, TestData)
```

Here, to normalize the dataset scale() function is used. Package neuralnet has a function neuralnet() to train the model. Function compute() is used for predicting the result.

$$\mathrm{y(x) = f(}\sum_{i=1}^{n} w\mathrm{i}x i) \tag{8}$$

where x_i are input variables, y is the output variable, and $\mathrm{w_i}$ are weights.

Activation function used is sigmoid.

$$\mathrm{A_j} = \frac{1}{1 + e^{-x}} \tag{9}$$

3.5 Results and Graphical Analysis

To evaluate the results and error rate, we have considered mainly six performance measures.

$$\text{MAE (Mean Absolute Error)} = \frac{1}{n}\sum_{i=1}^{n}\left|yi - \hat{y}\mathrm{i}\right| \tag{10}$$

It is the average absolute difference between your prediction and observed reality.

$$\text{MSE (Mean Squared Error)} = \frac{1}{n}\sum_{i=1}^{n}\left(yi - \hat{y}\text{i}\right)^2 \tag{11}$$

$$\text{RMSE (Root Mean Squared Error)} = \sqrt{\text{MSE}} \tag{12}$$

$$\text{MAPE (Mean Absolute Percentage Error)} = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{yi - \hat{y}\text{i}}{y\text{i}}\right| \tag{13}$$

$$\text{R} - \text{squared} = \frac{\text{Explained variation}}{\text{Total variation}} \tag{14}$$

R^2 value always lies between 0 and 1. COR is used here to find the Pearson correlation coefficient. Its value ranges from −1 to +1. Higher the R^2 and correlation value, better the model fits the data. Hence, the values nearer to 1 mean model is more efficient.

COR and R^2 values of SVR are 0.93 and 0.86, respectively. The values given by backpropagation neural network are 0.90 and 0.81, respectively, and that of multiple linear regressions are 0.87 and 0.75, respectively. Both the values are nearer to 1 in SVR compared to the other two models. Therefore, SVR is yielding better prediction results than the other two models. Next to SVR, values of backpropagation neural network are better compared to multiple linear regressions [7].

A graph of actual and predicted values is plotted for each of the mobile phones [8]. The blue color indicates the actual price, and red indicates the predicted line. Overlapping points denote that actual and predicted prices are almost the same. As the gap between lines increases, the difference between actual and predicted values also increases. In SVR, lines are overlapping more than remaining two (Figs. 3, 4, and 5).

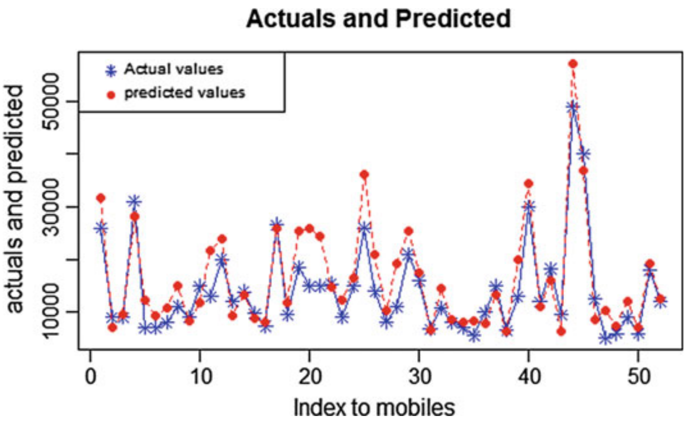

Fig. 3 Support vector regression—price versus index to mobiles

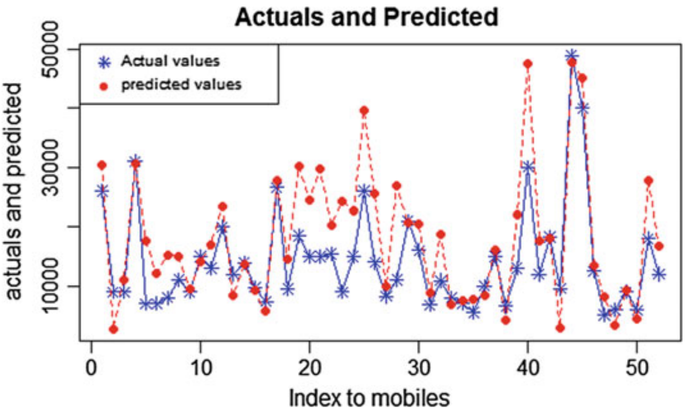

Fig. 4 Multiple linear regressions—price versus index to mobiles

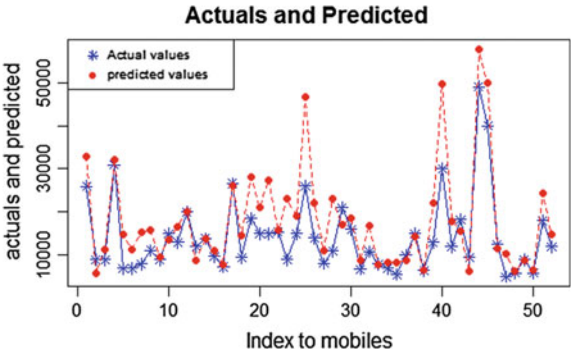

Fig. 5 Backpropagation neural network—price versus index to mobiles

A graph of actual versus predicted is plotted for all three models. Points on the line denote predicted values equal to the actual values. Points below the line indicate predicted values are less than the actual price. Points above the line indicate predicted price is more than the actual price. Hence, the model with more points on or nearer to the line gives a better prediction. Here, comparing three graphs SVR is having more points nearer to the line (Figs. 6, 7, and 8).

Comparing graphs and performance measures of all three models, we can arrive at a conclusion that SVR is yielding more efficient results than the other two models.

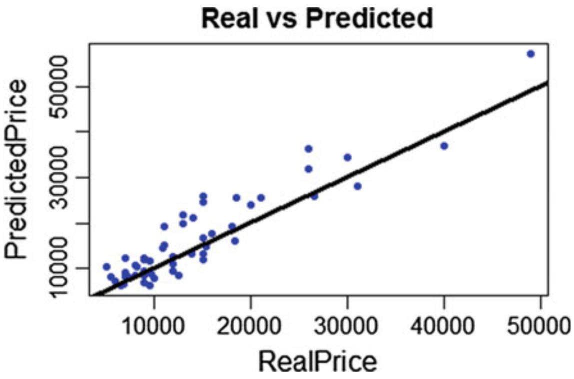

Fig. 6 Support vector regression—actual versus predicted

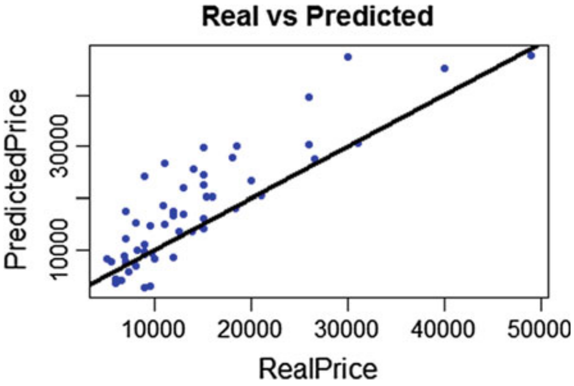

Fig. 7 Multiple linear regressions—actual versus predicted

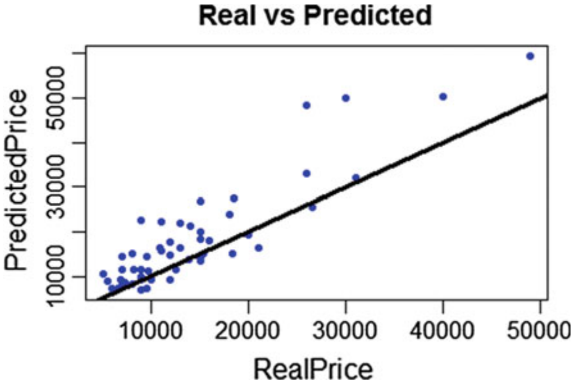

Fig. 8 Backpropagation neural network—actual versus predicted

4 Conclusion

In this paper, we have discussed three machine learning regression models for the prediction of price. The main aim of proposing these three models is to find out which algorithm gives more accurate results for price prediction. The prediction is done by considering different features of the smartphones which are preprocessed, and then the data are given for prediction. From the results, we can say that SVR algorithm is giving more accurate results which is then followed by backpropagation neural networks and at last multiple linear regressions. Hence, SVR is the best-suited algorithm for good price prediction. These models can be used in the retail industry for pricing their products.

5 Future Scope

In the future, we can extend these predictive models for prediction of the price of other products like laptops and camera. It can also be used with few modifications as required for other structured and unstructured datasets such as images. The dataset can be stored in Hadoop Distributed File System (HDFS) to enhance scalability of the data volume; we can easily store a huge volume of dataset required for prediction.

References

1. Kalaiselvi N, Aravind KR, Balaguru S, Vijayaragul V (2017) Retail price analytics using back-propagation neural network and sentimental analytics
2. Hegadi RS et al (2013) Statistical data quality model for data migration business enterprise. Int J Soft Comput 8:340–351. https://doi.org/10.3923/ijscomp.2013.340.351
3. Chandrashekhara KT et al (2015) Complex event processing in smart homes. Int J Sci Eng Appl Sci
4. Manjunath TN et al (2013) Data quality assessment model for data migration business enterprise. Int J Eng Technol (IJET) 5(1). ISSN 0975-4024
5. Chen C-C, Kuo C, Kuo S-Y, Chou YH (2015) Dynamic normalization BPN for stock price forecasting. In: International conference on systems, man, and cybernetics
6. Meesad P, Rasel RI (2013) Predicting stock market price using support vector regression
7. Pushpa SK, Manjunath TN, Mrunal TV, Singh A, Suhas C (2017) Class result prediction using machine learning. In: 2017 international conference on smart technologies for smart nation (SmartTechCon), Bangalore, 2017, pp 1208–1212
8. Gireesh Babu CN et al (2017) Real-time data processing with storm: using twitter streaming. https://doi.org/10.5281/zenodo.822928