Price Forecasting for Products in E-Commerce using Learning Algorithms

VIJAY KUMAR PALADUGU[1], DR. M. RAMA BAI[2]

Mahatma Gandhi Institute of Technology (A),Hyderabad ,India.

Abstract

E-Commerce websites offer a variety of offers throughout the year, where people end up buying the product less likely for a fair price or a slight loss or great loss. This proposed study analyzes and builds Supervised Machine Learning models that forecast a product's price at a given timestamp.The work made use of models like Support Vector Regression (SVR), Random Forest Regression, Decision Tree Regression and Gradient Boosting Regression, which are regression-based algorithms that are trained and tested on the Flip Kart dataset for price prediction given the product details with a time stamp. The dataset consists of Twenty Thousand records out of which more than Five Thousand missing data is observed in the brand feature and treated by imputation. The study concludes that Random Forest model is able to explain the Variance Score more by 0.67 than other three models. Hence Random Forest model was able to capture more variability between the multiple independent features and dependent feature.

Keywords: **Price Prediction, Random Forest, Regression.**

1. INTRODUCTION

Through E-commerce almost anything can be purchased today. It is conducted over computers, tablets, smartphones, and other smart devices. The three areas of e-commerce are online retailing, electronic markets, and online auctions, which is supported by electronic business. The existence of E-commerce not only allow consumers to order online but also to pay online through the Internet to save the time and money of consumers and enterprises thereby it also greatly improved transaction efficiency today, especially for busy office workers, it saves a lot of valuable time. In the present problem statement Price prediction, Forecasting factor like price, a continuous feature has become important because the price factor can't be known until the day of purchase and which always will be changing, in other scenarios it's difficult to buy the product because the product has become slightly or more expensive than earlier though the product is required within a short period. Flip Kart products dataset provide details like the time stamp, product details like name, retail price, and discounted price. Price prediction is a regression task and can be handled through Supervised Learning based techniques.

In Ahmed Fathalla,Ahmad Salah, Kenli Li, Keqin Li and Piccialli Francesco [1] proposed to predict the price of second hand items using SVR and BI-LSTM-CNN, among which BI-LSTM-CNN model made good performance (77.14) compared to SVR (66.43) and MAE for BI-LSTM-CNN AND SVR are 0.0702 and 0.0929 respectively. Though in this work the prediction was made on second hand items wherein the algorithms ran to learn the behavior of owners of second items but current research was made on new first hand items using machine learning Algorithms. According to Jianhu Zheng, Mingfang Huang [2] proposed work it proved that Long Short Term Memory (LSTM) was able outperform than other classic machine learning algorithms like ARIMA and Back Propagation neural network (BPNN). In [2] LSTM algorithm was able give relatively good prediction when compare to other algorithms and the RMSE score is 61.16, 26.87, 14.44 for ARIMA, BPNN and LSTM respectively. In Qian Yunneng [3], the work takes the historical information of stock as input and make prediction. In [3], early study which has one day previous stock information and has got the Standard Error as 3.97 and the proposed work had improved the KNN and Standard Error received is 3.66.In Chandrashekhara K.T, Thungamani M, Gireesh Babu C.N, Manjunath T.N [4] work for price prediction problem algorithms used are Multiple Linear Regression, Neural Network and SVR, among these SVR gave good performance which is measured by the metric called R Squared and the value is 86 percentage. In Kunal Pahwa,

Neha Agarwal [5], the work helps to make someone who want to use machine learning as their technique to build predictive model, it explains about various technique Algorithmic Assumptions, Hyper Parameter knowledge, Data Leakage and others.

In Susmita Ray [6],The work attempted to throw light on the Machine Learning algorithms merits and demerits based on their application, which can give an idea to appropriately selecting an algorithm according for the problem at hand.Author took various algorithms namely Gradient Descent Algorithm, Linear Regression Algorithm, Multivariate Regression Analysis, Logistic Regression, Decision Tree, Support Vector Machine, Bayesian Learning, Naive Bayes, K Nearest Neighbor Algorithm, K Means Clustering Algorithm, and Back Propagation Algorithm. Author say's Gradient Descent Algorithm is an iterative method in which the objective is to minimize cost function, there are three different types of Gradient Descent namely Stochastic Gradient Descent (SGD), Batch Gradient Descent (BGD), and Mini Batch Gradient Descent (MBGD). BGD is Computational efficient, it produces stable error gradient and stable convergence. It needs the entire data that needs to be trained be available in memory. In SGD error is calculated for each training example and parameters are updated, Advantage with this approach is that "Rate of Improvement" in the result can be seen on frequent updates but computational expensive. In MBGD the data is split into small batches and parameters are updated for these batches. Algorithm like Linear Regression attempts to fit a straight line on the data, this algorithm can be appropriate for the cases where the data has a linear relation, that means the covariate's and response variable has linear relation between them. In Multivariate Regression Analysis the optimal scenario is, that the covariates should be correlated with response variable but with each other. When there is potential correlation between the covariates among themselves then it is said to be multi collinearity. Complexity of this technique is high as it requires people to acquire knowledge and expertise on statistical techniques and statistical modeling. On the other hand algorithm like K Nearest Neighbor can be implemented easily and model building happens to be cheap. Its superiority is, it handle's the multi classification problem well and also it is a lazy learner because it computes the distance over K neighbors. In the prediction for protein function it performed better than SVM.

In Sayavoung Lounaapha, Wu Zhongdong , Chalita Sookasame [7],The work focused on stock price analysis. In particular, the study used the daily historical data set of the stock prices taken from the Stock Exchange of Thailand (SET). The Convolutional Neural Network (CNN) method had been fairly fit to the data and the performance is high in accuracy, CNN is type of feed-forward network which means to say that the connection of input layer to hidden layer and then hidden layer to output layer are in the forward fashion and hence, this study also proves that CNN can be used for similar kind of problems with good performance.In Manas Ranjan Senapati, Sumanjit Das, Sarojananda Mishra [8],The work attempted in predicting the stock price using Hybrid Neural Network, in the earlier attempts author's which have worked on same stock price prediction have commonly applied Artificial Neural Network (ANN) technique, also the result of using Neural Network outshine than Normal Feed-Forward Neural Network in estimating the stock market growth. In this paper, the algorithm used is Ada Line Network which went through optimization by modified PSO to predict the open stock price, also it was compared with other two-hybrid technique. The proposed model has lesser Mean Absolute Percentage Error than other schemes. In Chenhao Wang, Qiang Gao [9], The work predict the Soya bean prices in the future, unlike the closing price which earlier research was done instead the study concentrated in the high and low price factors for the stock which can give better results. This study used deep learning technique for the prediction of Soya bean future prices in particular it took LSTM neural network and BP Neural Network and. In Naalla Vineeth, Maturi Ayyappa, B Bharathi [10],The work uses a Neural Network model in predicting the house price. The dataset has 19 house sales features of the king country, USA. The study initially started with Simple Linear Regression for building the model, then applied Multiple Linear Regression and also used Neural Network. All these methods went through evaluation by the metric called "Root Mean Squared Error" (talks about the differences between actual values and predicted values). The results at the end turn to be favor to Neural Networks with an error of 2.190. In Abirami R, Vijaya MS

[11], This paper attempted to predict future stock price by eroding a historical dataset collected from Aditya Birla money Ltd. The dataset consists of 970 instances and 7 attributes: opening price, previous closing price, highest price, lowest price, last price, average price, and closing price. Models used are SVR and Linear regression and among them SVR yield good performance in particular kernel parameter is set to Radial Bias Function. Software used was WEKA and the evaluation metrics and their respective results for SVR are, for Mean Squared Error it is 0.55132 and accuracy is 94 percentage.

2. METHODOLOGY

2.1 System Architecture

The work proposes the System architecture of choosing the best performing algorithm. Firstly, Collect the data and pre-process it to make it free from missing values and transform the text data into numerical values then we go for data analyzing which is a crucial step in order to get the insights in the data and also which help to make the data perfect or digestible form such that the results acquired at last are reliable.
Before training and testing begin the whole data is split into two half's the first half is training data usually considered to be 70 or 75 percent of the total data and second half is test data which is the leftover 30 or 25 percent data from the total data and the model builds on the training data and then tested on the testing data. Then accuracy of the algorithm is been noted and further repeate the whole process from training till noting down accuracy for all regression-based algorithms. Finally the model that achieved higher accuracy is selected.

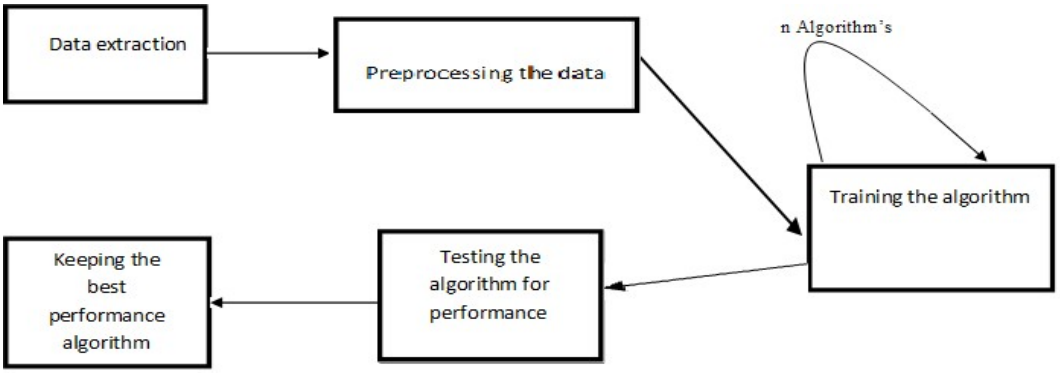

Figure1: System Architecture

2.2 Feature Engineering

Feature exploration is done for all the expected predictors and found that Feature extraction and feature Imputing methods can be applied which can help preparing the data for the model training.

2.2.1 Feature Extraction: The mid function in excel generates the "Category" feature from the feature "Product Category Tree".

Table 1: Illustration of category feature extraction

Product category tree	Category
["Clothing >> Women's Clothing >> Lingerie, Sleep & Swimwear >> Shorts >> Alisha Shorts >> Alisha Solid Women's Cycling Shorts"]	Clo
["Furniture >> Living Room Furniture >> Sofa Beds & Futons >> Fab Home Decor Fabric Double Sofa Bed (Finish Colo..."]	Fur
["Footwear >> Women's Footwear >> Ballerinas >> AW Bellies"]	Foo
["Clothing >> Women's Clothing >> Lingerie, Sleep & Swimwear >> Shorts >> Alisha Shorts >> Alisha Solid Women's Cycling Shorts"]	Clo
["Pet Supplies >> Grooming >> Skin & Coat Care >> Shampoo >> Sicons All Purpose Arnica Dog Shampoo (500 ml)"]	Pet
["Eternal Gandhi Super Series Crystal Paper Weight..."]	Ete
["Clothing >> Women's Clothing >> Lingerie, Sleep & Swimwear >> Shorts >> Alisha Shorts >> Alisha Solid Women's Cycling Shorts"]	Clo

From the feature Pid (Sub category feature) is generated using left function in excel.

Table 2: Illustration of sub category feature extraction

Pid	Sub category
SRTEH2FF9KEDEFGF	SITE
SBEEH3QGU7MFYJFY	SEE
SHOEH4GRSUBJGZXE	SHOE
SRTEH2F6HUZMQ6SJ	SITE
PSOEH3ZYDMSYARJ5	POSE
PWTEB7H2E4KCYUE3	PWTE
SRTEH2FVVKRBAXHB	SITE

2.2.2 Filling the missing data: Filling the blank or missing or NAN (not a number) values is crucial as it can miss leads the algorithm in learning wrong things, as concerned to brand variable there are more than 5000 which are missing. Below discussed two techniques to treat missing values.

- Dropping the missing data:
  As there are many missing data, it could be a loss to drop the missing values as they can hold valuable information.

- Imputing the missing data:
  In the study it would be great loss of information by dropping the missing values so we opt imputing as an option. Since the missing data is categorical, we used the left () in excel to fetch the starting few characters of product name as it can hold the brand information.

Table 3. Illustration of actual brand feature and new brand feature extracted from product name feature

Product name	Brand (Actual)	Brand (with imputed values)
Sicons Conditioning Conditioner Dog Shampoo	Sicons	Sicons
dongle Printed Boy's Round Neck T-Shirt	Dongle	Dongle
SWAGGA Women Clogs	SWAGGA	SWAGGA
Kennel RubberDumbell With Bell - Small Rubber Rubber Toy For Dog	Kennel	Kennel
Glass WeddingLingerie Set	NAN	Glass Wedding
Veelys Shiny White Quad Roller Skates - Size 4.5 UK	NAN	Veelys Shiny
Bulky vanity case Jewellery Vanity Case	NAN	Bulky vanit

3. EXPERIMENTAL RESULTS

3.1 Filp kart Dataset

The Data set is an open-source provided by Kaggle which consists of 20,000 records or rows from which 81 records ar missing which can be removed as its percentage is 0.405 to overall records.

3.2 SVR (Support Vector Regression)

SVR provides kernels namely 'rbf', 'linear','poly' and for our problem we used 'rbf'. Other parameters that are taken for building the model are C = 40 and gamma = 4.Both of these parameters are the optimal team for the problem.
In this problem the prediction is made on the “offer” feature which is extracted from the feature retail price and discount price.

Table 4: Illustration of SVR Results

Index	Actual	prediction
18205	91.8197	73.72543252
6950	19.9	40.45055184
4265	59.91984	65.56248067
10108	92	34.27573976
856	33.33333	76.87380521
7051	83.8785	85.75426112
3616	71.42857	32.92691771

3.3 Gradient Boosting Regression algorithm

The Gradient boosting regression algorithm used for regression problems takes the decision tree's average performance, for which the applied learning rate is 1.0.In Gradient boosting regression algorithm as the trees get increased the performance goes better which directs us towards over fit so we took just learning rate as hyper parameter and adjusted to 1.

Table 5. Illustration of Gradient Boosting Regression Results

Index	Actual	Prediction
18205	91.8197	80.70557
6950	19.9	35.15215
4265	59.91984	58.60214
10108	92	55.21147
856	33.33333	47.44764
7051	83.8785	89.26341
3616	71.42857	62.68453

3.4 Decision Tree Regression algorithm

Decision Tree Regression is used for regression problems, in the decision tree as the no of trees are increased, there is a high risk of over fitting. In the proposed work the default parameters as is are used.

Table 6: Illustration of Decision Tree Regression Results

Index	Actual	Prediction
18205	91.8197	59.91984
6950	19.9	20.65
4265	59.91984	49.0982
10108	92	67.72727
856	33.33333	49.96094
7051	83.8785	95.11002
3616	71.42857	71.42857

3.5 Random Forest Regression algorithm

Unlike other algorithms, random forest is not prone to over fit as the trees increase, and the hyper parameter selection is done using the Randomized Search CV.
The final optimal parameter which the work used are as follows:
N Estimators =50, Min Samples split = 6, Min Samples Leaf=1, Max Features = 'Square Root', Ma x Depth=40, and Bootstrap=False.

Table 7: Illustration of Random Forest Regression Results

Index	Actual	Prediction
18205	91.8197	79.01506
6950	19.9	20.66543
4265	59.91984	58.6567
10108	92	55.51658
856	33.33333	53.2086
7051	83.8785	87.81112
3616	71.42857	72.53515

4. CONCLUSION AND DISCUSSION

Each model made an effort to learn the relationship between independent variables and dependent variable which contributed in forecasting the price. All of the algorithms learned on the same data but Random Forest stood outperform, the remaining algorithms due to their built-in or architectural limitations could not do better as Decision Tree Regression, Gradient Boosting Regression has the over fitting risk while SVR has Hyper parameter such as C and Gamma parameters which help it not to fall in the trap but even that could not yield model to perform better. The work witnessed that any regression problem at hand Random Forest Regression not only does the best job in predicting but also finishes training fast. Random Forest Regression obtained 0.67 variability of different Brands.
Further study can be taken up by including more information on the products that has its records in different time steps, which will help the models to learn more and improve the performance.

REFERENCES

[1] Ahmed Fathalla, K. L. K. L. P. F., Ahmad Salah. Deep end-to-end learning for price prediction of second-hand items. Knowledge and InformationSystems, July 2020.

[2] Jianhu Zheng, M. H. Traffic flow forecasting through time series analysis based on deep learning. IEEE, 2020.

[3] Yunneng, Q. A new stock price prediction model based on improved knn. International Conference on Information Science and Control Engineering(ICISCE), (7), July 2020.

[4] Chandrashekhara K.T, G. B. C. M. T., Thungamani M. Smartphone price prediction in retail industry using machine learning techniques. Emerging Research in Electronics, Computer Science and Technology, 545, 2019.

[5] Kunal Pahwa, N. A. Stock market analysis usingsupervised machine learning. IEEE.

[6] Ray, S. A quick review of machine learning al- gorithms. International Conference on MachineLearning, Big Data, Cloud and Parallel Computing Engineering, Februry 2019.

[7] Sayavoung Lounaapha, C. S., Wu Zhongdong. Research on stock price prediction method based on convolutional neural network. International Conference on Virtual Reality and Intelligent Systems (ICVRIS), 2019.

[8] Manas Ranjan Senapati, S. M., Sumanjit Das. A novel model for stock price prediction using hybrid neural network.Journal of the Institution ofEngineers, June 2018.

[9] Chenhao Wang, Q. G. High and low prices pre- dictions of soya bean futures with lstm neural net- work. International Conference on Software Engineering and Service Science (ICSESS), 2020.

[10] Naalla Vineeth, B. B., Maturi Ayyappa. House price prediction using machine learning algorithms. [Communication in Computer and Information Science]Soft Computing Systems, 837, 2018.

[11] Abirami R, V. M. Stock price prediction using support vector regression. Global Trends in Computing and Communication Systems, 269, 2012.