MI Based Mobile Price Analyzer for Indian Market

Bhanu Jaisri. R. G[1], Dr. K. Kanagaraj[2]

[1]Student, Department Computer Applications, Mepco Schlenk Engineering College, Sivakasi
[2]Associate Professor, Department of Computer Applications, Mepco Schlenk Engineering College, Sivakasi.

Abstract

The mobile market in India grows rapidly due to more inflow of new designs with advanced features and mindset of customers to switch over from one brand / model to other for grabbing advantages of additional special features. Price monitoring and prediction tool is essential to capture different types of Indian customers against mobiles in Indian markets. Both Beautifulsoup and Selenium web scraping methods are utilized to fetch required data from the pre-defined commercial websites. Data preprocessing is carried out before applying algorithms.

Three prediction models with Random Forest, XGBoost and Gradient Boosting are designed to get accurate price prediction of customer chosen mobile among the existing Indian sellers. The developed models are trained with 80% of extracted data and remaining 20% is used for testing and evaluation. The performance of each model is tested, evaluated and compared with metrics. The analysis of test results proved that GB model has gained the highest accuracy with R-square value 0.9829 and values of RMSE, MAE and MSE are with n limits.

Keywords: Web Scraping, Beautifulsoup, Selenium, Data Preprocessing, Prediction Accuracy

INTRODUCTION

Price Prediction plays a dominant role in E-marketing / e-commerce platform as it is a base line for any business model to gain enough profit margin to sustain in the volatile competitive market. Designing a price prediction tool suitable for the subject under study and acquisition of relevant data set are critical due to complex correlation among impact factors. Accounting more factors may confuse customers in making conclusive decision for purchasing the desired product of their choice. But, considering few factors to develop price predictive model yields false prediction. Hence, selection of optimum influential factors through knowledge / experience on e-commerce and price prediction analysis is essential.

The prime goal of any business promoting software is providing user friendly comfort platform to customer such that he can choose product of his interest easily including brand, model number, features and comparative statement of techno-commercial aspects of different retailers. In global market, India is the main focus of all mobile manufacturers as it has better scope of market than other countries. The scope of this research is limited to mobiles available in Indian market to meet Indian customer requirements. The proposed software aids end customer to select any brand of mobiles with inherent specification launched in Indian local / online markets. It provides comparison of price quoted by all Indian E-Commerce platforms / online retailers like Amazon, Flipkart, Indiamart, Jiomart and so on, against the

model number and brand of mobile finalized by customer. Different algorithms used for price prediction related to several products are

- Random Forest Regressor (RF)
- Gradient Boosting Regressor (GB)
- XGBoost Regressor (XGB)

Periodically updated dataset yields better accurate result than old data. Data pre-processing is mandatory before feeding dataset as input to any algorithm. This paper compares the results of three algorithms with current data set.

LITERATURE SURVEY

The accuracy of any developed prediction model depends not only on design parameters but also on input dataset. Due to volatile nature of digital market particularly for mobiles, effective price prediction within time schedule is mandatory to gain advantage over other similar products and attract end customer. Generally, RF algorithm is preferred for price prediction of any fast-moving product due to its simplicity and ease of use. Machine Learning (ML) based house price prediction model using RF Regression is demonstrated with 70% of collected data for training and balance 30% for testing [1].

XGB, an algorithm under gradient boosting framework offers high efficiency, flexibility and portability in predictive analysis. It outperforms than other algorithms in terms of accuracy [2]. In gold price prediction, empirical results obtained from forecast model designed with XGB algorithm reveal better performance than CatBoost, LightGBM [3].

In case of larger data, GB is preferred than RF Regression to achieve highest possible accuracy. Mostly in stock market, GB is used to predict forecast data very close to futuristic real data [4]. Prediction model for urban house price is designed to facilitate real estate companies and capital investors to arrive conclusive decision on purchase / investment / contract negotiation. Three ML algorithms namely Linear Regression, RF, and GB are employed on Melbourne housing dataset for prediction analysis. The Performance of any developed model is evaluated using metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and Coefficient of Determination (R2 Score). Among the models tested, GB emerges as the most effective model [5].

Web Scraping is an automated systematic sequential approach to acquire useful valid information from e-commerce websites [6]. The first step in web scraping is applying Beautifulsoup to scrape website data related to subject of interest. BeautifulSoup uses HTML parser from Python's standard library. Process begins with Fetching Phase, where relevant data is accessed from predefined website using HTTP protocol. Finally, it stores extracted data in the desired format. In few cases, it fails due to crawling problems. Selenium is an active web scraping technique which opens the selected websites by its own to solve crawling problem. Thus, web scraping extracts structured data from HTML text element [7]. The seamless flow built in web scraping technique ensures extraction of data from web pages and retrieval of data from storage for further analysis and processing [8].

RESEARCH METHODOLOGY

The objective of this research is developing three mobile price prediction models independently by using RF, XGB, GB to suit needs of Indian customer. The scope of work is broadly classified into three area.

1. Data Collection
2. Data Pre-Processing

3. Training, Testing and evaluation of model

1. Data Collection

In the present study, data collection is a process of acquiring inherent data of mobiles in stock or having short delivery period in Indian local/retail/digital market like brand, model, specifications, actual cost, discount prices and the source URL. The existing kaggle dataset does not contain current information pertaining to each brand of mobile. Hence, using this data as input for price prediction may not yield fruitful result. In this paper, data collection methodology shown in Fig 1 applies adaptive technique for acquisition of current data against updated information in websites. Web scraping with Beautiful Soup and Selenium is used to extract selective information from e-commerce websites. Irrespective of Beautifulsoup and selenium, web scraping sends request to the desired URL and it receives HTML document as response. These tools extract structured data from websites. The developed logic directs extracted data to store in Azure database for analysis.
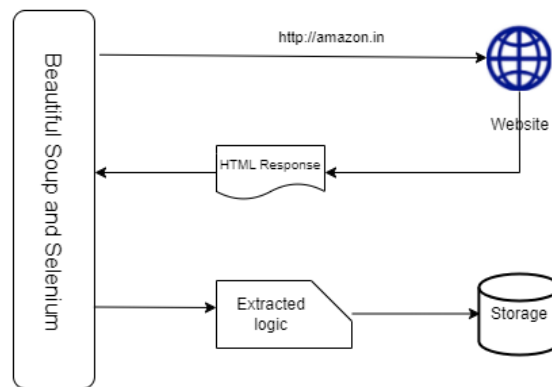

Fig 1 Workflow of Web Scraping

2. Data Pre-processing

Data pre-processing is a reliable process of ensuring integrity of acquired data by removing duplicate information and adding missing details related to aimed objectives. It helps to minimize errors in data processing and predictive analysis. Further, it normalizes the numerical values of dominant price prediction factors. In the present analysis, heatmap is used to determine the most relevant features of the selected mobiles through correlation analysis.

3. Training, Testing and Evaluation of model

Three price prediction models have been built using RF, XGB and GB respectively to achieve accurate prediction to promote mobile business in Indian dynamic market. From the selected data domain, 80% of pre-processed data is allotted for traing of each model. During training, each selected model is verified against its intended function. Training is a prerequisite for testing and evaluation of any model. In test phase, balance 20% valid data serve as input to evaluate performace metrics corresponding to each model. Finally, the achieved results against each model are compared to decide best model for the selected application and dataset.

RESULTS AND DISCUSSION

The price prediction of new generation mobiles posted in different websites by e-commerce retailers is critical due to volatile nature of digital mobile market and dynamic behaviour of Indian customers. Due to non availability of current valid dataset suitable for the selected application, web scraping tools like beautifulsoup and selenium are used to extract required data. The extracted dataset is pre-processed to

meet input requirements of the proposed model. Three regression techniques RF, XGB and GB are deployed to develop mobile price prediction models to suit the selected application.

RF Regression model is tested and performance metrics like MAE, MSE, RMSE and R-Squared are evaluated. The achieved results of RF model are listed in Table 1. From the analysis of RF model results, the numerical values of indices such as MAE, RMSE, MSE are nominal. The RF model has achieved R-Square of 0.9817.

TABLE 1. RF TEST RESULTS

Performance Metrics	Achieved Numeric Values
MAE	781.51
MSE	8065765.29
RMSE	2840.03
R-Squared	0.9817

XGB model utilizes parallel tree mechanism to achieve accurate results in faster mode. The evaluated performance indices are tabulated in Table 2. The scores of MAE, MSE, RMSE are within acceptable limits and accuracy is 0.9812.

TABLE 2. XGB TEST RESULTS

Performance Metrics	Achieved Numeric Values
MAE	804.20
MSE	8310055.69
RMSE	2882.72
R-Squared	0.9812

GB model uses ensemble technique which contains number of decision trees to attack regression problems. It builds series of decision trees and operates sequentially to minimize prediction errors. The performance metrics obtained during testing of GB model is displayed in Table 3. The GB model has achieved highest accuracy of 0.9829.

TABLE 3. GB TEST RESULTS

Performance Metrics	Achieved Numeric Values
MAE	781.39
MSE	7567508.31
RMSE	2750.91
R-Squared	0.9829

TABLE 4. COMPARE THREE MODELS

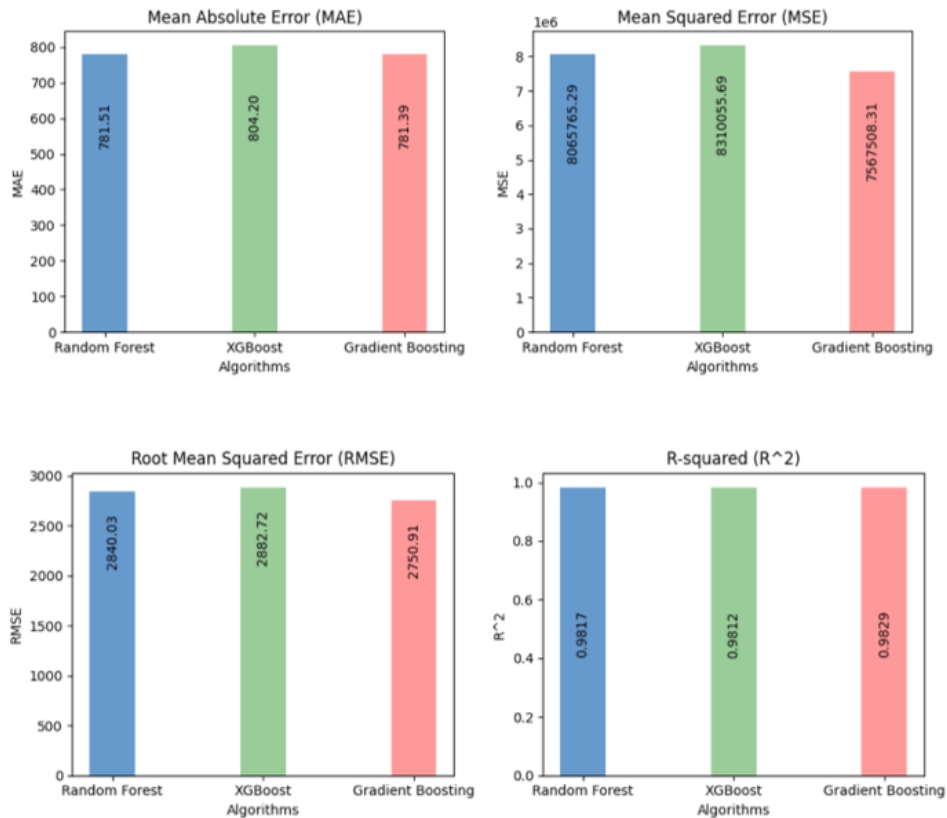| Algorithm	MAE	MSE	RMSE	R-Squared
RF	781.51	8065765.29	2840.03	0.9817
XGR	804.20	8310055.69	2882.72	0.9812
GB	781.39	7567508.31	2750.91	0.9829

Fig 2. Algorithm Comparison

The test results of all three models are compiled as shown in Table 4 and Fig 2 for comparison. Among the selected three models, GB model outperforms by producing better accuracy with acceptable scores of other metrics.

CONCLUSION

The input dataset required for training, testing and evaluation of price prediction analyzer suitable for all brands of mobiles launched in Indian market, is extracted using web scrap method. Three models are developed to resolve short comings of mobile price prediction prevailing in Indian volatile market. All three models are successfully trained, tested and evaluated against desired functions and performance metrics. The analysis of test results reveals that GB model is better and most suitable for predicting mobile price of Indian customer's choice. The present study can be extended to global customers.

REFERENCES

1. Adetunji, A. B., Akande, O. N., Ajala, F. A., Oyewo, O., Akande, Y. F., & Oluwadara, G. (2022). House Price Prediction using Random Forest Machine Learning Technique. Procedia Computer Science. 199, 806–813. DOI: 10.1016/j.procs.2022.01.100.
2. Gono, D. N., Napitupulu, H., & Firdaniza(2023). Silver Price Forecasting Using Extreme Gradient Boosting (XGBoost) Method. Journal of Mathematical Forecasting, 10(3), 150-165. DOI: 10.3390/math11183813.
3. Ben Jabeur, S., Mefteh-Wali, S., & Viviani, J.-L. (2021). Forecasting gold price with the XGBoost algorithm and SHAP interaction values. Annals of Operations Research. DOI: 10.1007/s10479-021-04187-w

4. Shrivastav, L. K., & Kumar, R. (2022). An Ensemble of Random Forest Gradient Boosting Machine and Deep Learning Methods for Stock Price Prediction. Journal of Information Technology Research, 15(1), 1-19. IGI Global. DOI: 10.4018/JITR.2022010102

5. EL Mouna, L., Silkan, H., Haynf, Y., Nann, M. F., & Tekouabou, S. C. K. (2023). A Comparative Study of Urban House Price Prediction using Machine Learning Algorithms. E3S Web of Conferences, 418, 03001. DOI: 10.1051/e3sconf/202341803001

6. Zhao, B. (2017). Web Scraping. Springer International Publishing (Vol. 567, pp. 483-491). DOI: 10.1007/978-3-319-32001-4_483-1.

7. Singh, S., Shukla, A., Devanshu, & Gandhi, A. B. (2022). Web Scraping Using Selenium. International Journal of New Research in Discipline, 7(1), ISSN: 2456-4184. Paper Id: IJNRD2201018

8. Rajeswari, M., & Pavyaasree, D. (2022). Scrape the Product Details from Ecommerce Websites. Journal of Emerging Technologies and Innovative Research (JETIR), 9(6), 158. ISSN: 2349-5162. Paper ID: JETIRFN06026.

9. Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. METHODS ARTICLE. DOI: 10.3389/fnbot.2013.00021.

10. Song, P.Y, & Liu, Y.T (2021). An XGBoost Algorithm for Predicting Purchasing Behaviour on E-Commerce Platforms. Theoretical Economics Letters. DOI: 10.17559/TV-20200808113807.

11. Nabi, R. M., & Saeed, S., (2020). A Novel Approach for Stock Price Prediction Using Gradient Boosting Machine with Feature Engineering. Kurdistan Journal of Applied Research, 5(2), 87-94. DOI: 10.24017/science.2020.1.3.

12. Almaslukh, B. (2020). A Gradient Boosting Method for Effective Prediction of Housing Prices in Complex Real Estate Systems. 2020 International Conference on Technologies and Applications of Artificial Intelligence. DOI: 10.1109/TAAI51410.2020.0004.

13. Deshmane, V., Musale, J., Joshi, S., Chinta, V., Gokak, K., & Dalbhanjan, I. (2023). Web Scraping for E-Commerce Website. IJIEMR Transactions, 13(04), 216-224. DOI: 10.48047/IJIEMR/V13/ISSUE04/24.