Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/48/20241516

Mobile phone price prediction: A comparative study among four models

Qipeng Liang

School of mathematical and physical sciences, University of Nanjing Tech, Nanjing, China

202121144047@njtech.edu.cn

Abstract. As science and technology is advancing by leaps and bounds, mobile phones have become part and parcel of people's life. Because the different models of mobile phones which have different structural foundations, the prices of mobile phones are constantly fluctuating. Mobile phone prices forecasts are becoming more precise as artificial intelligence develops. This article compares various machine learning approaches, and the importance of the variables is ranked in order to determine the most accurate way to forecast the prices of mobile phones. The machine learning techniques used are linear regression (LR), random forest regressor (RFR), XGB Regressor and Support Vector Machine regressor (SVM). In order to determine which model predicts the most accurate mobile phone prices, R^2 evaluation is used. The XGB Regressor model had the greatest score (R-squared = 0.95) for prediction of mobile phone prices, compared to the other three models. In a word, with XGB Regressor methodology as a priority for future mobile phone price predicting, which can improve the accuracy of price predicting.

Keywords: Price prediction, Machine learning techniques, XGB Regressor.

1. Introduction

The era of the smartphone has arrived. Mobile phones are becoming increasingly powerful in functionality, people rely on them more than ever before, and people have become inseparable from mobile phones. The price of mobile phones has also had a crucial impact on people's lives, and they are now more inclined to invest in mobile phones with higher cost performance. Therefore, accurate prediction of the price of mobile phones can produce huge benefits for both buyers and sellers.

The prices of mobile phones can be accurately predicted utilizing machine learning algorithms. In this research, linear regression, random forest regressor (RFR), XGB Regressor (XGBoost) and Support Vector Machine regressor (SVM) are used to train a more accurate model. The primary objective of this research is to train the model using these variables in order to identify the model that achieves the highest accuracy in predicting true values and evaluate the model with R-squared evaluation. In the results section, the chart of the differences between the actual price and predicted price of the train set and the test set, the checking residuals chart of the training set, ranking chart of the importance of impact factors and the R-squared evaluation table are drawn.

DOI: 10.54254/2755-2721/48/20241516

2. Literature Review

The mobile phone has become an indispensable aspect of human life, and the utilization of appropriate algorithms enables the prediction of mobile phone prices. This not only assists customers in purchasing more cost-effective devices but also serves as a valuable reference for mobile phone manufacturers [1]. In fact, extensive research has been conducted on the prediction of mobile phone prices. It is possible to forecast the price scope of a mobile phone by classifying the many factors that influence it. The machine learning algorithm demonstrates exceptional accuracy in accurately forecasting the price scope of mobile phones [2]. When predicting mobile phone prices, a diverse range of variables related to mobile phones are utilized to optimize the impact factor and achieve a predicted mobile phone price that aligns with real-world observations [3]. Employing supervised machine learning techniques in the domain of machine learning to train models and generate predictions, thereby determining the most suitable algorithm for the given dataset to make phone price predictions [4]. By employing the linear regression methodology to establish a more accurate model for mobile phone price prediction, so that people can better predict the price of mobile phones to choose more cost-effective mobile phones [5]. When employing machine learning techniques for predicting future mobile phone prices, linear regression and XGB Regressor methodologies can be utilized to construct models that enhance the accuracy of mobile phone price predictions [6]. In order to investigate the ranking of impact factors affecting the importance of mobile phone prices, investors can employ the variable influence method in random forest regressor. Additionally, within the context of cooperative games, examining the relative contributions of different variables can enhance the precision in predicting mobile phone prices [7]. The price of a mobile phone can be predicted employing a random forest machine learning technique, which enables parameter pruning in the model and consequently enhances the accuracy and precision of the decision tree, bring benefits for prediction of mobile phone prices [8]. Models were constructed using XGB Regressor and Support Vector Machine methods, utilizing historical data on mobile phone price, which helps to improve the accuracy of predicting mobile phone prices [9]. Due to the lack of available resources for cross-checking phone prices, individuals are unable to make informed decisions when purchasing a mobile phone. Therefore, employing Support Vector Machine as a training model can enhance the accuracy of mobile phone price prediction and purchase more suitable mobile phone [10]. However, current investigations are still limited and does not reach an agreement on the predictions of mobile phone prices. Thus, in-depth investigations are needed.

3. Data

3.1. Dataset

The dataset utilized in this research is sourced from kaggle, including a multitude of factors pertaining to mobile phones. The dataset consists of 161 rows and 14 columns, wherein each row represents a distinct phone model along with its corresponding impact factors on price. The 14 columns are displayed in Table 1.

Table 1. Presentation of the dataset

Column	Description
1st column	ID of each cellphone
2nd column	Price of each cellphone
3rd column	Sales number of each cellphone
4th column	Weight of each cellphone
5th column	Resolution of each cellphone
6th column	Phone Pixel Density of each cellphone
7th column	Type of CPU core in each cellphone
8th column	CPU Frequency in each cellphone

DOI: 10.54254/2755-2721/48/20241516

Table 1. (continued).

9th column	Internal memory of each cellphone
10th column	Random Access Memory of each cellphone
11th column	Number of Rear Cameras of each cellphone
12th column	Number of Front Cameras of each cellphone
13th column	Battery Capacity of each cellphone
14th column	Thickness of each cellphone

3.2. Data Splitting

Before splitting, it is required to find any missing values and remove such data from the dataset. After data cleaning, basic information is depicted in Table 2. In addition, the train set, and the test set are separated, accounting for 70% and 30% of the total dataset. There are 112 rows of data for the training set and the left 49 rows for the testing set. Training data (112 rows): training model. Testing data (49 rows): testing model.

Table 2. Statistical presentation of variables

Variable	Count	Mean	Std	Min	Max
Product_id	161.0	675.6	410.9	10.0	1339.0
Price	161.0	2215.6	768.2	614.0	4361.0
Sale	161.0	621.5	1546.6	10.0	9807.0
Weight	161.0	170.4	92.9	66.0	753.0
Resolution	161.0	5.2	1.5	1.4	12.2
Ppi	161.0	335.1	134.8	121.0	806.0
Cpu core	161.0	4.9	2.4	0.0	8.0
Cpu freq	161.0	1.5	0.6	0.0	2.7
Internal mem	161.0	24.5	28.8	0.0	128.0
Ram	161.0	2.2	1.6	0.0	6.0
Rearcam	161.0	10.4	6.2	0.0	20.0
Front_cam	161.0	4.5	4.3	0.0	20.0
battery	161.0	2842.1	1367.0	800.0	9500.0
thickness	161.0	8.9	2.2	5.1	18.5

3.3. Data Multicollinearity

It is necessary to conduct correlation analysis on the dataset. Through multicollinearity analysis, it is found that the value of the multicollinearity between weight and battery is 0.8, and the value of the multicollinearity between resolution and battery is 0.8, indicating that these two variables are highly correlated with the battery, so this research dropped the weight and resolution, which makes the dataset more representative.

4. Methodology

Among all machine learning methodologies, the linear regression, random forest regressor (RFR), XGB Regressor (XGBoost), and Support Vector Machine regressor (SVM) were selected. These four regression methodologies are utilized for model training to compare the actual price with the predicted price to evaluate the goodness of fit of model. The models are then evaluated using R-squared evaluation to determine the most suitable one for the dataset.

DOI: 10.54254/2755-2721/48/20241516

4.1. Linear Regression

Linear regression is a machine learning technique that utilizes mathematical statistics to establish the quantitative correlation among multiple variables. It employs regressor analysis to approximate the connection establishing a linear relationship between a single independent variable and a corresponding dependent variable, known as unitary linear regression analysis. On the other hand, if multiple independent variables are involved in the analysis and there is a direct correlation between the dependent variable and multiple independent variables, it is known as an analysis technique called multiple linear regression. In conclusion, linear regression trains a model based on the given training data and uses this model to make predictions, assuming that the characteristic satisfies the linear relationship. Linear regression plays a fundamental role in statistical modeling [11].

4.2. Random Forest Regressor

The Random Forest Regressor is a machine learning technique that utilizes ensemble learning techniques, a powerful technique in which diverse algorithms or the same algorithm can be combined multiple times to construct an enhanced predictive model. Specifically, random forests amalgamate multiple decision trees, which enhances their collective efficacy. Random forest regressor can also achieve good performance by using larger primary tuning parameter values [12].

4.3. XGB Regressor

The XGB Regressor belongs to the class of Boosting algorithms, which aims to combine multiple weak classifiers into a powerful classifier. As XGB Regressor is based on an ensemble of decision trees, it effectively integrates numerous trees models a strong classifier. The XGB Regressor is acknowledged for its remarkable predictive capability as an algorithm [13].

4.4. Support Vector Machine Regressor

Support Vector Machine regressor is to find the best line or the best hyperplane to classify the data. In a word, finding the classification hyperplane and the maximizing the class spacing to achieve classification. Ultimately, Support Vector Machine regressor becomes an optimization problem of maximizing the interval. The Support Vector Machine Regressor is demonstrated as a valuable tool for multivariate calibration, particularly in cases where outliers and non-linearities are present [14].

4.5. Evaluation Parameters

To assess the goodness of fit of the four models, two evaluation methodologies were employed in this study: R-squared value and mean absolute error (MAE).

5. Experimental results

The dataset for this study comes from kaggle. Four machine learning methodologies are used in this study: linear regression, random forest regressor (RFR), XGB Regressor (XGBoost) and Support Vector Machine regressor (SVM), using training dataset to train models and using testing dataset to predict phone prices. In addition, the importance ranking of the impact factors in random forest regressor is also used to rank the importance of the impact factors affecting the price of mobile phones. The related results are shown in Figures 1-5.

DOI: 10.54254/2755-2721/48/20241516

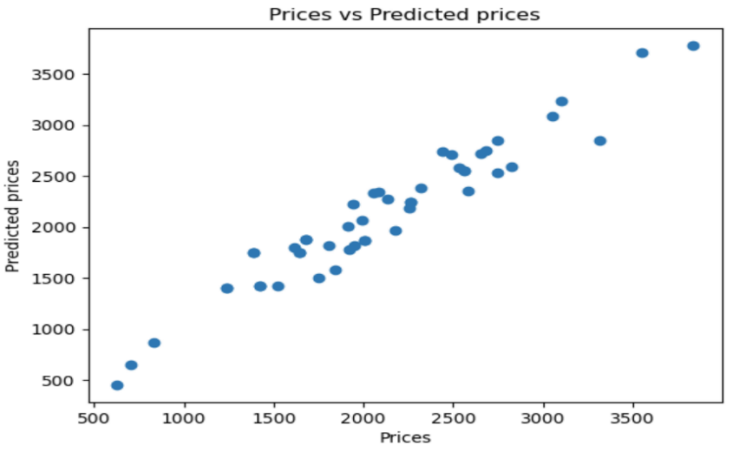

Figure 1. Real price and the forecast price under the linear regression model.

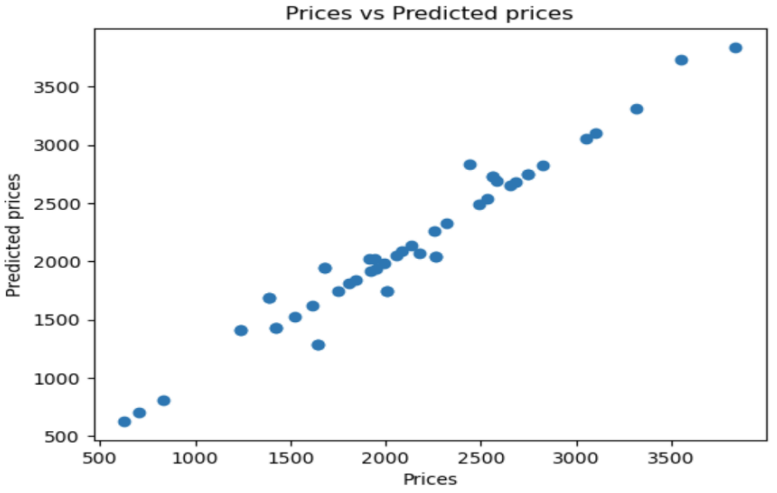

Figure 2. Real price and the forecast price under the random forest regressor model.

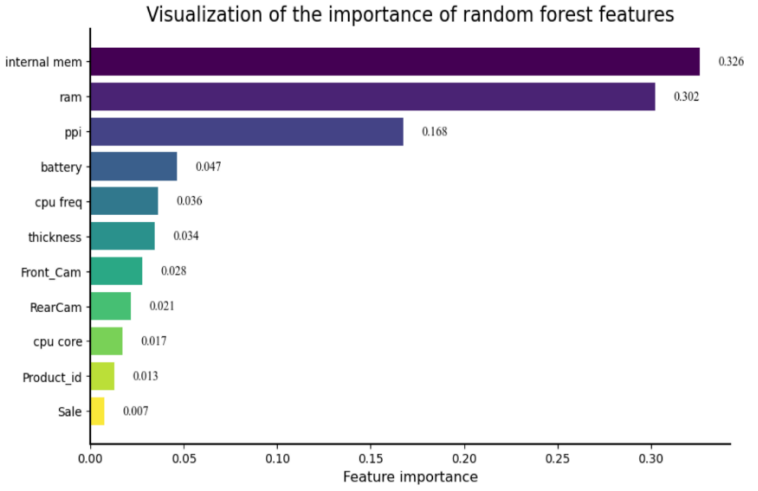

Figure 3. Importance ranking under the random forest regressor model.

DOI: 10.54254/2755-2721/48/20241516

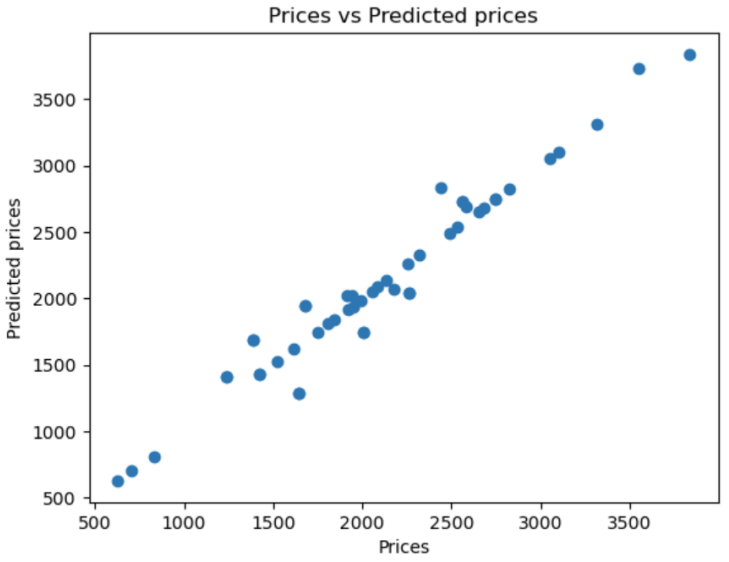

Figure 4. Real price and the forecast price under the XGB Regressor model.

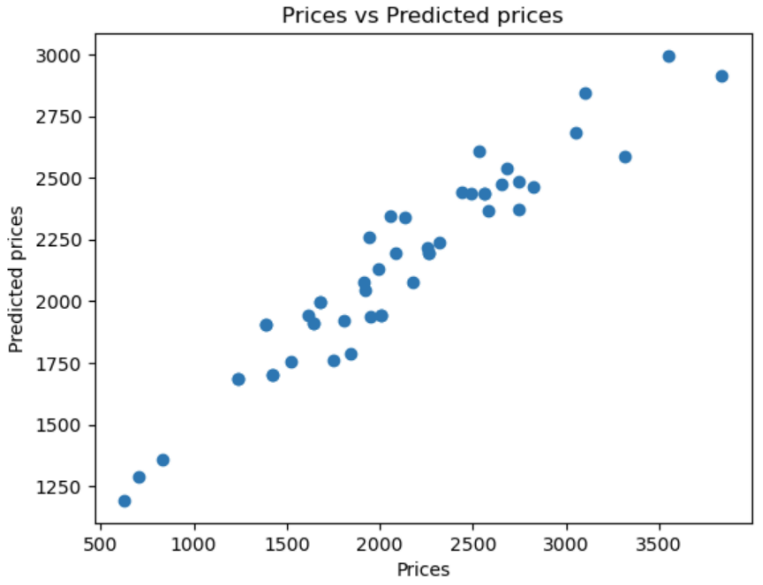

Figure 5. Real price and the forecast price under the Support Vector Machine regressor model.

The following Table 3 shows the comparison between the models. Obviously, the XGBoost performs the best.

Table 3. R-squared comparison between models

Model	R-squared Score
XGBoost	0.949136
Random Forest	0.939511
Linear Regression	0.929672
Support Vector Machine	0.767690

DOI: 10.54254/2755-2721/48/20241516

6. Conclusion

This study uses four regression models, including LR, RFR, XGB Regressor and SVM regressor to train the model of training set, followed by evaluation using R-squared on the testing set. Obviously, XGB Regressor exhibited the highest score (R-squared = 0.95), indicating its superior goodness of fit for this particular dataset than other three regressor methodologies. Furthermore, this it can be concluded that the internal memory is the most significant factor for pricing the mobile phone. However, limitations still exist, for example, more machine and more sophisticated models should be introduced in this field for predicting the prices of mobile phones.

References

[1] Subhiksha, S., Thota, S., & Sangeetha, J. (2020). Prediction of phone prices using machine learning techniques. In Data Engineering and Communication Technology: Proceedings of 3rd ICDECT-2K19 (pp. 781-789). Springer Singapore.

[2] Nasser, I. M., & Al-Shawwa, M. (2019). ANN for predicting mobile phone price range. Int J Acad Inf Syst Res (IJAISR), 3(2).

[3] Chandrashekhara, K. T., Thungamani, M., Gireesh Babu, C. N., & Manjunath, T. N. (2019). Smartphone price prediction in retail industry using machine learning techniques. In Emerging Research in Electronics, Computer Science and Technology: Proceedings of International Conference, ICERECT 2018 (pp. 363-373). Springer Singapore.

[4] Kiran, A. V., & Jebakumar, R. (2022). Prediction of Mobile Phone Price Class using Supervised Machine Learning Techniques. International Journal of Innovative Science and Research Technology, 7, 248-251.

[5] Rani, S., Kumar, S., Jain, A., & Swathi, A. (2022, October). Commodities Price Prediction using Various ML Techniques. In 2022 2nd International Conference on Technological Advancements in Computational Sciences (ICTACS) (pp. 277-282). IEEE.

[6] Fofanah, A. J. (2021). Machine learning model approaches for price prediction in coffee market using linear regression, XGB, and LSTM techniques. International Journal of Scientific Research in Science and Technology, (6).

[7] Hur, J. H., Ihm, S. Y., & Park, Y. H. (2017). A variable impacts measurement in random forest for mobile cloud computing. Wireless communications and mobile computing, 2017.

[8] Sakib, A. H., Shakir, A. K., Sutradhar, S., Saleh, M. A., Akram, W., & Biplop, K. B. M. B. (2022, January). A hybrid model for predicting Mobile Price Range using machine learning techniques. In 2022 The 8th International Conference on Computing and Data Engineering (pp. 86-91).

[9] Jose, J., Raj, V., Seaban, S. V., & Jose, D. V. (2023, February). Machine Learning Algorithms for Prediction of Mobile Phone Prices. In International Conference On Innovative Computing And Communication (pp. 81-89). Singapore: Springer Nature Singapore.

[10] Kalaivani, K. S., Priyadharshini, N., Nivedhashri, S., & Nandhini, R. (2021, November). Predicting the price range of mobile phones using machine learning techniques. In AIP Conference Proceedings (Vol. 2387, No. 1). AIP Publishing.

[11] Su, X., Yan, X., & Tsai, C. L. (2012). Linear regression. Wiley Interdisciplinary Reviews: Computational Statistics, 4(3), 275-294.

[12] Segal, M. R. (2004). Machine learning benchmarks and random forest regression.

[13] Pesantez-Narvaez, J., Guillen, M., & Alcañiz, M. (2019). Predicting motor insurance claims using telematics data—XGBoost versus logistic regression. Risks, 7(2), 70.

[14] Brereton, R. G., & Lloyd, G. R. (2010). Support vector machines for classification and regression. Analyst, 135(2), 230-267.