International Journal of Applied Engineering & Technology

Mobile Phone Price Prediction Based on Supervised Learning Algorithms

Aries Maesya

Universitas Pakuan, a.maesya@unpak.ac.id

Yanfi Yanfi

Bina Nusantara University, eufrasia.yan.fi@binus.ac.id

Lukas

Universitas Katolik Indonesia Atma Jaya, lukas@atmajaya.ac.id

Date of Submission: 22nd December 2022 Revised: 8th January 2023 Accepted: 26th January 2023

How to Cite: **Maesya, Aries et. al., (2023). Mobile Phone Price Prediction Based on Supervised Learning Algorithms. International Journal of Applied Engineering and Technology 5(1), pp 41-44.**

***Abstract* - The price of mobile phones is one of the most important factors in the success of a mobile product in the market. Regression methods to predict the price of mobile phones based on its features can help companies to decide the price of a new mobile phone. This study investigates the variables that significantly predict the price and develop the models to predict the price using two methods, which are linear regression and random forest methods. The experiment uses data downloaded from Kaggle containing 145 mobile phone prices and features. It is found that the linear regression and random forest algorithms can provide a relatively good prediction of mobile phones with MAPE scores below 10% and R2 scores above 95%. The random forest method predicts the price slightly better than linear regression.**

***Index Terms* - mobile phone pricing prediction, linear regression, random forest, mean absolute percentage error, root mean square error, coefficient of determination.**

INTRODUCTION

Mobile phone has turned into an indispensable item for many people, where 91% of the people in the world own mobile phones and 83.3% of the people specifically own smartphones today [1]. There are more than 170 mobile phone brands originating from 38 countries, competing in national and international markets [2]. Market acceptance of a mobile phone depend on many factors, such as the price, features, shape, and brand. In such a competitive market, price can be very significant in determining the success of a mobile phone.

There are some studies that predict mobile phone price range using machine learning, such as using neural network [3], K-nearest neighbors [4], [5], deep neural network [4], random forest [5], logistic regression [5], decision tree [5], linear discriminant analysis [5], and support vector [5]. These studies predict the price classes of mobile phones, not the actual price.

Haomeng Liu [6] developed an automatic pricing for refurbished mobile phones using a fuzzy neural network coupled with a momentum technique for parameters optimization (FNN-MUP). The study focused on 12 variables that are considered significant to influence the price of refurbished mobile phones, such as: brand, model, technical specifications, and the states of the mobile phone.

There is still a lack of study on the model of mobile phone pricing model, therefore this study investigates which features of mobile phone significantly affect its price and predict the price of mobile phone based on the features. Two algorithms are compared in this study, namely linear regression, and random forest algorithms with the following research questions:

RQ1: What are the features of mobile phone that significantly influence the price of the mobile phone?

RQ2: How accurate are linear regression and random forest methods in predicting the market value of mobile phones?

The random forest [7]–[9] is a popular learning method in the category of supervised learning, that is improved over the decision tree learning approach [10]. The learning method can be used for prediction by averaging the results of multiple decision trees with a random selection of features. It has been successfully used for price prediction of the house [11], stock [12], cloud service [13], and many more.

Another very popular supervised learning approach is the linear regression. The regression learning approach predicts the target value by creating a linear function of the input values. It is also commonly used to predict prices such as the price of houses [14], [15], stock [16], gold [17], and others.

Random forest and linear regression are used in this study for the prediction of mobile phone prices using the data sample of 145 mobile phone prices along with their features, which are obtained from Kaggle [18].

The accuracy of the two models is measured in terms of Root Mean Squared Error (RMSE), coefficient of determination (R2), and Mean Absolute Percentage Error (MAPE).

The resulting experiments show that random forest regression and linear regression algorithms can provide a good prediction of mobile phone price with relatively small errors, where MAPE scores are below 10% and R2 scores above 95%. It is found that the random forest approach is marginally better than the linear regression approach in predicting the price.

METHOD

This research uses a Supervised Learning Algorithm with a forecasting method by utilizing Linear Regression and Random Forest Regression algorithms with two distribution of data scenarios, namely 80:20 and 70:30 to determine the best algorithm performance. This research method has detailed steps as presented in Figure 1.

I. Data Acquisition

This study uses data downloaded from Kaggle. The data obtained is mobile phone price prediction with a total of 2268 data. The variables in the data source include product id, price of the mobile phone, quantity of sale, weight, display resolution, number of pixels per inch, number of CPU Cores, CPU frequency, size of internal memory, size of RAM, resolution of rear camera, resolution of front camera, battery capacity and phone thickness. Example of the dataset is depicted in Table I.

II. Data Selection

The data selection process is the process of selecting relevant data or variables for this research. The variables in the data source include Product_id, Price, Sale, Weight, Resolution, PPI, CPU Core, CPU Freq, Internal memory, RAM, Rear Cam, Front Cam, Battery, and Thickness.

III. Data Cleaning

In data cleaning after data selection, the data is cleaned of data redundancy (duplicate data) and missing variable data to remove empty data and clean up unnecessary syntax.

IV. Data Transformation

The data transformation process is the process of combining data into an appropriate form. In this process, numeric data will be converted into a value range of 0-1 for all attributes. Data transformation is used to process or adjust the data so that we can effectively employ Linear Regression and also Random Forest Regression algorithm.

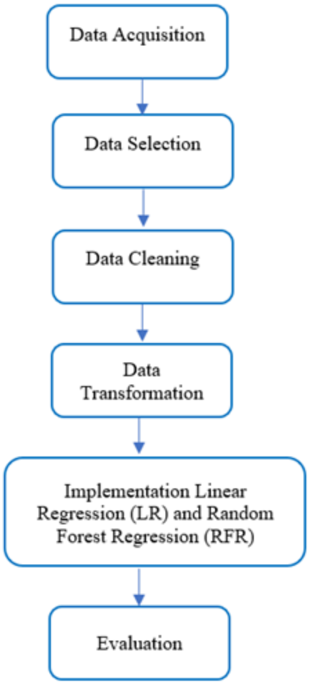
Data Acquisition
Data Selection
Data Cleaning
Data Transformation
Implementation Linear Regression (LR) and Random Forest Regression (RFR)
Evaluation

FIGURE 1 RESEARCH METHOD

TABLE I
SAMPLE OF DATA SOURCE VALUE

Product_id	Price	Sale	...	Front_Cam	battery	thickness
203	2357	10	...	8.0	2610	7.4
880	1749	10	...	0.0	1700	9.9
40	1916	10	...	5.0	2000	7.6
99	1315	11	...	0.0	1400	11.0
880	1749	11	...	0.0	1700	9.9

V. Random Forest Regression

Random Forest Regressors are often used to solve problems related to classification, regression and so on. This algorithm is a combination of several tree predictions or Decision Tree Regressors where the prediction results from Random Forest are obtained through the most results from each individual decision tree [19], so for Random Forest it is formulated as follows:

$$l(y) = argmax_c \sum_{n=1}^{N} I_{h_n y(c)} \qquad (1)$$

Where:

I = indicator function
hn = number of trees in Random Forest

VI. Linear Regression

This study also performs multiple-variable regression. The battery, CPU Frequency, CPU Core, Front Camera, Internal memory, Phone Pixel Density (ppi), RAM, Rear Camera, resolution, sales number, thickness, and weight are identified as predictors, while price is the response.

Thus, the formula is described as follows:

$$Price = a + bX \qquad (2)$$

Where:
Price = variable response
a = constant
b = regression coefficient
X = predictor variables

```
%MULTIPLE VARIABLE REGRESSION
X = [battery, CPU_Freq, CPU_Core, Front_Camera,
internalMem, ppi, ram, RearCam, resolution, Sale,
thickness, weight]

%Fit linear model regression
model = fitlm (X, Price)
```

In order to improve accuracy for low to medium-dimensional data sets, we utilize "fitlm" to fit the linear regression model.

VII. Evaluation Model

Model evaluation in regression is different from classification. In the search regression model, the evaluation of the model sought is the error value in a model. evaluation of the regression model is carried out in several ways, including:

a. *RMSE:* evaluating the model by looking at the level of error in the prediction results.

b. *R^2:* evaluating the proportion of variables in data that is calculated to see the fit of a model.

c. *MAPE:* statistically evaluating the accuracy of the predictions in the forecasting method, by comparing the forecast error and the actual value.

Result And Discussion

I. Data Analysis

Before starting the model process, the raw data with unnecessary variables such as Product_id and sale are temporarily dropped, and records with a missing value are also deleted.

II. Data Visualization

This stage is used to see the correlation between the Price variable and the others. This results in the plot that can be seen in Figure 2.

III. Training Data and Test Data

For training and evaluating the model, we split the data into training and test data with a proportion of 80:20 and 70:30 for comparison.

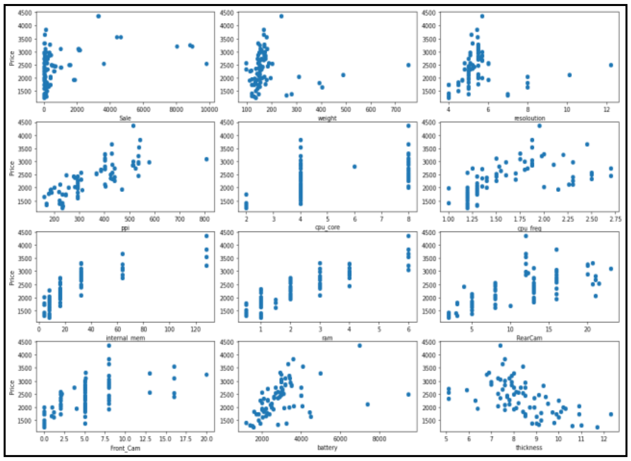

FIGURE 2 Correlation Between Variables

IV. Correlation Analysis

Correlation analysis is a technique to determine if a linear relationship exists between pair of variables. When a relationship exists between two variables, then changes in one variables (X) will cause corresponding changes in the other variable (Y). This is depicted in Figure 3.

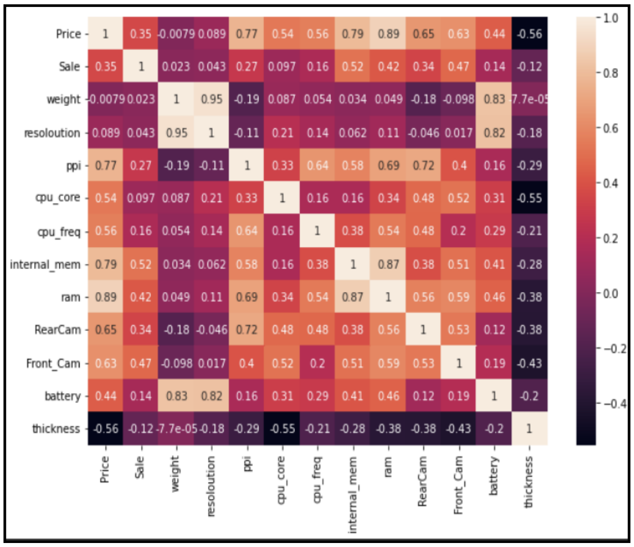

FIGURE 3 Correlation of Mobile Phone Price

V. Density Estimation

At this stage do the distribution of the probability between the actual values and Prediction values of the variable Price. The results is presented in Figure 4.

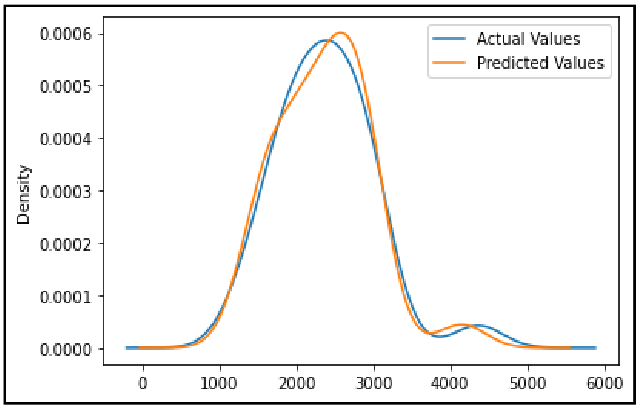

FIGURE 4 VISUALIZATION OF KERNEL DENSITY ESTIMATION

VI. Linear Regression

Table II described that resolution and weight are not significantly correlated with the Price. However, the most correlated with Price is RAM, follows by internal memory, phone pixel density (ppi), rear camera, front camera, CPU frequency, thickness, CPU core, and battery.

TABLE II
LINEAR REGRESSION WITH A SINGLE VARIABLE TO PRICE

Variables	**R2**	**P-Value**
Battery	0.194	0.000
CPU Core	0.288	0.000
CPU Frequency	0.311	0.000
Front camera	0.395	0.000
Internal Memory	0.627	0.000
Phone Pixel Density (ppi)	0.590	0.000
RAM	0.785	0.000
Rear Camera	0.425	0.000
Resolution	0.008	0.289
Thickness	0.309	0.000
Weight	0.000	0.925

Linear regression with multiple variables is also conducted. The result of the estimated coefficients is displayed in Table III. The independent variable (X) consists of twelve variables along with one dependent variable (Price), namely: battery, CPU Frequency, CPU Core, Front Camera, Internal Memory, ppi, ram, Rear Camera, resolution, thickness, and weight. Moreover, it proves that the front camera, rear camera, and weight are not significantly correlated to the price.

In addition, as the resolution and thickness variable decrease, the price tends to increase.

We found the R2 value to be larger than 0.9, and the p-value is much less than the default significance level. We can conclude that a significant linear regression relationship does exist between the response price variable and the predictor variables in X.

TABLE III
ESTIMATED COEFFICIENTS OF MOBILE PHONE PRICE

Variables	**Estimate**	**SE**	**tStat**	**pValue**
(Intercept)	2197.200	266.440	8.247	0.000
Battery	0.145	0.031	4.675	0.000
CPU Frequency	121.250	48.582	2.496	0.014
CPU Core	51.170	10.346	4.946	0.000
Front Camera	7.137	5.109	1.397	0.165
Internal Memory	5.112	1.172	4.363	0.000
Phone Pixel Density (ppi)	1.081	0.225	4.814	0.000
RAM	96.088	26.395	3.640	0.000
Rear Camera	2.697	4.330	0.623	0.534
Resolution	-149.270	54.411	-2.743	0.007
Thickness	-94.354	14.868	-6.346	0.000
Weight	0.264	0.810	0.325	0.745

VII. Evaluation Model

The testing modeling in this research uses a partitioning modeling design with train data and test data of 80:20 and 70:30. Then, prediction testing is conducted using a Random Forest regressor resulting in a score of 95.53% while prediction testing using Linear Regression resulting in a score of 95.11% as tabulated in Table IV.

TABLE IV
COMPARISON OF ACCURACY MODELS

Method	**Accuracy Model**	
	80:20	**70:30**
Random Forest Regression	95.55%	94.29%
Linear Regression	95.11%	95.54%

TABLE V
Comparison Of Regression Model Evaluation

Method	**RMSE**		**MAPE**		**R^2**	
	80:20	**70:30**	**80:20**	**70:30**	**80:20**	**70:30**
Random Forest Regression	148,73	179.50	6.10%	7.31%	95.53%	94.29%
Linear Regression	155.50	158.67	6.91%	6.58%	95.11%	95.54%

Based on the results of the study as tabulated in Table V, it was found that the accuracy value that had been tested was partitioning with training and testing data of 80:20 and 70:30.

In testing portioning training data was 80% and testing data 20%, the best accuracy value is Random Forest Regression of 95.55%, while portioning training data of 70% and testing data of 30 produces the best accuracy value, namely Linear Regression of 95.54%.

Based on the research results obtained, it can be concluded that the two prediction methods that have been used to predict mobile phone prices, namely Random Forest Regression and Linear Regression are able to provide a good error score with MAPE scores below 10% and r2 scores above 94%. Between the two models, the Random Forest Regression model got better prediction results with an R2 Score of 95.53%, a MAPE Score of 6.1%, and an RMSE Score of 148,733. Because in Random Forest regression there are parameters that determine the number of decisions to be used.

Conclusion

Based on the research results obtained, it can be concluded that from both single and multiple-variable linear regression, resolution and weight are not significantly related to Price. Besides, the two prediction methods that have been used to predict mobile phone prices, namely Random Forest Regression and Linear Regression can provide a good error score with MAPE scores below 10% and R2 scores above 94%.

Between the two models, the Random Forest Regression model got better prediction results with an R2 Score of 95.53%, a MAPE Score of 6.1%, and an RMSE Score of 148,733. Because in Random Forest Regression there are parameters that determine the number of decisions to be used.

For further research, we suggest adding more additional data sources and trying different supervised learning methods.

References

[1] Bankmycell.com, “How many smartphones are in the world?” https://www.bankmycell.com/blog/how-many-phones-are-in-the-world (accessed Oct. 09, 2022).

[2] “List of mobile phone brands by country,” Wikipedia. https://en.wikipedia.org/wiki/List_of_mobile_phone_brands_by_country (accessed Oct. 09, 2022).

[3] I. M. Nasser and M. Al-Shawwa, “Developing Artificial Neural Network for Predicting Mobile Phone Price Range,” International Journal of Academic Information Systems Research, vol. 3, no. 2, 2019.

[4] E. Güvenç, G. Çetin, and H. Koçak, “Comparison of KNN and DNN Classifiers Performance in Predicting Mobile Phone Price Ranges,” Advances in Artificial Intelligence Research (AAIR), vol. 1, no. 1, 2021.

[5] M. Cetin and Y. Koc, “Mobile Phone Price Class Prediction Using Different Classification Algorithms with Feature Selection and Parameter Optimization,” 2021. doi: 10.1109/ISMSIT52890.2021.9604550.

[6] H. Liu, J. Huang, H. Han, and H. Yang, “An Improved Intelligent Pricing Model for Recycled Mobile Phones,” 2020. doi: 10.1109/CAC51589.2020.9327611.

[7] T. K. Ho, “Random decision forests,” in Proceedings of the International Conference on Document Analysis and Recognition, ICDAR, 1995, vol. 1. doi: 10.1109/ICDAR.1995.598994.

[8] T. K. Ho, “The random subspace method for constructing decision forests,” IEEE Trans Pattern Anal Mach Intell, vol. 20, no. 8, 1998, doi: 10.1109/34.709601.

[9] L. Breiman, “Random forests,” Mach Learn, vol. 45, no. 1, 2001, doi: 10.1023/A:1010933404324.

[10] L. Rokach and O. Maimon, Data mining with decision trees: Theory and applications. 2007. doi: 10.1142/6604.

[11] A. B. Adetunji, O. N. Akande, F. A. Ajala, O. Oyewo, Y. F. Akande, and G. Oluwadara, “House Price Prediction using Random Forest Machine Learning Technique,” in Procedia Computer Science, 2021, vol. 199. doi: 10.1016/j.procs.2022.01.100.

[12] P. Sadorsky, “A Random Forests Approach to Predicting Clean Energy Stock Prices,” Journal of Risk and Financial Management, vol. 14, no. 2, 2021, doi: 10.3390/jrfm14020048.

[13] V. Khandelwal, A. K. Chaturvedi, and C. P. Gupta, “Amazon EC2 Spot Price Prediction Using Regression Random Forests,” IEEE Transactions on Cloud Computing, vol. 8, no. 1, 2020, doi: 10.1109/TCC.2017.2780159.

[14] A. Kaushal and A. Shankar, “House Price Prediction Using Multiple Linear Regression,” SSRN Electronic Journal, 2021, doi: 10.2139/ssrn.3833734.

[15] Q. Zhang, “Housing Price Prediction Based on Multiple Linear Regression,” Sci Program, vol. 2021, 2021, doi: 10.1155/2021/7678931.

[16] A. Gupta and T. J. Nagalakshmi, “Stock price prediction using linear regression in machine learning,” International Journal of Innovative Technology and Exploring Engineering, vol. 8, no. 12, 2019, doi: 10.35940/ijitee.L3932.1081219.

[17] I. Priyadi, J. Santony, and J. Na’am, “Data Mining Predictive Modeling for Prediction of Gold Prices Based on Dollar Exchange Rates, Bi Rates and World Crude Oil Prices,” Indonesian Journal of Artificial Intelligence and Data Mining, vol. 2, no. 2, 2019, doi: 10.24014/ijaidm.v2i2.6864.

[18] F. v. Younesi, S. Sahoo, and W. Ribeiro, “Mobile Price Prediction,” Kaggle, 2022. https://www.kaggle.com/datasets/mohannapd/mobile-price-prediction (accessed Oct. 09, 2022).

[19] G. A. Sandag, “Prediksi Rating Aplikasi App Store Menggunakan Algoritma Random Forest,” CogITo Smart Journal, vol. 6, no. 2, 2020, doi: 10.31154/cogito.v6i2.270.167-178.

Author Information

Aries Maesya, Computer Science Department, Universitas Pakuan, Bogor, Indonesia, 16142.

Yanfi Yanfi, Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia, 11480.

Lukas, Cognitive Engineering Research Group (CERG), Universitas Katolik Indonesia Atma Jaya, Jakarta, Indonesia 12930