Smartphone hedonic price study based on online retail price in Indonesia

Smartphone hedonic price study based on online retail price in Indonesia

T Listianingrum*, D Jayanti, and F M Afendi

Department of Statistics, Bogor Agricultural University (IPB), Bogor, 16680, Indonesia

*E-mail : tri_li@apps.ipb.ac.id

Abstract. The price of a smartphone is determined by its series of specifications, which makes it suitable to be modelled with a hedonic approach. We collected smartphone prices from various e-commerce along with the specifications using web scraping method. This study demonstrates statistical analysis of smartphone prices in Indonesia by constructing a hedonic model and investigate key variables determining smartphone price. An OLS regression model is employed to total of 1120 handsets data to construct a hedonic model. In addition, random forest regression and XGBoost regression were used to determine the variable importance. The results indicated that only 2 of 5 brands are significantly influence the price of smartphones; Xiaomi and Realme. Meanwhile, the other variables such as RAM, Internal Storage, NFC, Screen Size and Weight have significant positive effect on smartphone prices. In contrast to the use of brand as a weighing basis for calculating CPI, we discovered that RAM is the ultimate key variable determining smartphone prices.

1. Introduction

According to the Cambridge dictionary, a smartphone is a mobile phone that can be used as a small computer and connects to the internet. It usually equipped with highly advanced features. A typical smartphone has a high-resolution touch screen display, Wi-Fi connectivity, Web browsing capabilities, and the ability to accept sophisticated applications. With the accelerating speed of technological advancement, smartphones have become a mandatory component of people' daily performance. Based on Statista data, smartphone users were estimated to reach 28% of Indonesia total population in 2019. It was rising 2% from the previous year and continue to increase for the following years. It is amongst commodities that taken into account for Consumer Price Index (CPI) compilation which determines inflation value.

Every smartphone comes with particular technical and performance attributes which are known as specifications. These make them differentiated products with many alternative designs and selling prices in the market. Consumers derive utility from the attributes of the product rather than the product itself [1]. It means, consumers do not purchase a smartphone as it is; but rather they buy bundles of its specifications and features that come with it. Smartphone price depends on the set of specifications embodied such as brand, battery duration, display size, weight, camera, radio, etc.

Hedonic price model decomposes the price of a product into respective components that determine the product price [2]. It can be applied to identify and measure the relationship between the specifications and the price of a smartphone. The basic approach in constructing hedonic models is using

linear regression with price as the dependent variable and attributes as independent variables. However, the functional form of hedonic models are varied, and include the classic linear feature model, the logarithmic price model, the semi-logarithmic price model, and the semi-parametric model [3]. Other hedonic models include the BOX-COX transform models, models with an interaction effect and models including nonlinear features.

The high penetration rate of internet in Indonesia has enabled e-commerce to develop rapidly. Online shopping became an unavoidable phenomenon. Digital gadgets, computers and accessories ranks fourth in the category of goods most often purchased online in Indonesia [4]. Due to shifted shopping pattern towards digital and for the feasibility of data collection, the online retail price used in the model. Web scraping method was applied to collect data of smartphones' prices and specifications.

Some prior studies on smartphone price analysis had conducted in several countries. Lin and Chen conducted a hedonic price analysis of mobile telephones for the Taiwan market in 2013 [5]. In 2019 Waseem Ahmad, Tanvir Ahmed and Bashir Ahmad studied the effect of smartphone attributes on their retail price using a log-linear hedonic price model in Pakistan [6]. Both studies used ordinary least square (OLS) regressions to construct hedonic models. However, market share for smartphones diverse between countries which make the price analysis conducted might not be applicable in Indonesia. Also, since smartphones market changes rapidly, a price study needs to be done to get the most recent examination.

Hedonic model based on OLS regression can be used to analyze and compare significant effects factors on prices. However, it could not show the degree of importance amongst variables. Gu and Xu studied the housing market hedonic price study in 2017 and employed boosting regression tree to analyze the variable importance and influence path of housing hedonic price. We applied two methods of machine learning regression: random forest regression and extreme gradient boosting (XGBoost) regression to construct hedonic models and analyze how variables importance vary among those. This paper is among the early research on hedonic analysis for smartphones in Indonesia. It aimed to fill this gap and to provide a formal statistical analysis of smartphone prices in Indonesia by constructing a hedonic model and investigate the key variables determining smartphones prices.

This research could provide manufacturers with information about how consumers are valuing different attributes of the smartphone, which will help them in developing innovation strategies. Manufacturers need to incorporate preferred specifications in their products to reduce the chances of failure. Introduction of the less desired product attributes could cause wastage of resources and loss to the society [7]. The result of this research could also provide input for CPI measurement. Currently, for calculating the CPI, the price data for smartphones is weighted based on the brand. This study could influence the selection variable used as a basis for weighting. The most important specifications on the model could also refer as a basis for price data collection.

2. Material and methods

2.1. *Data and variables*

Web scraping was the primary data collection method to gather the smartphone's price and characteristics from iprice.co.id. The data collected on January 1st 2020. Iprice is a price comparison website: it is collecting product information, including pricing, from participating retailers and then displaying that collection of information on a single results page in response to a search query. It also provides information about the product's attribute or characteristics. The scraper developed using Python programming language with the BeautifulSoup package. Five dominating brands of smartphones were being the target of scraping: Oppo, Xiaomi, Samsung, Vivo, and Realme with a total market share of 94% in Indonesia [8]. The scraped data stored in a CSV format.

The scraped data still contained a significant amount of missing values and encountered inconsistency on the specifications. It needed to be cleaned and validated before further processing. Information from www.gsmarena.com used as a reference for validating the data as well as filling the missing values on specifications. Gsmarena.com is known to be a reliable website as a reference for

phone information. It established in June 2000, and currently has more than 20 billion information pages on it.

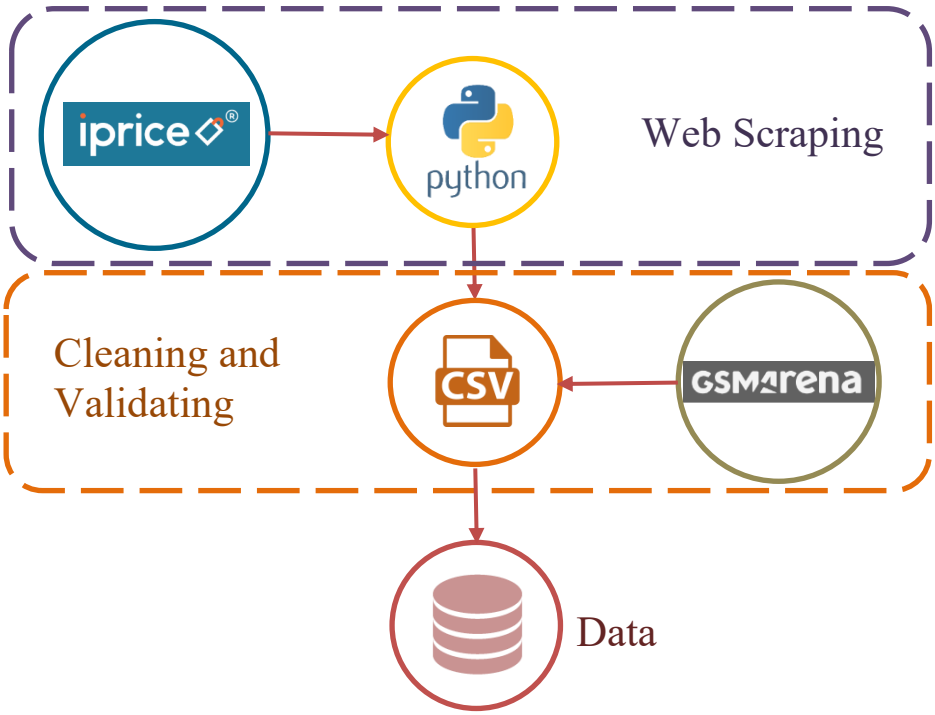

Figure 1. Data collection method

Web scraping generated 1286 rows of handset prices and characteristics. For each handset, prices information gathered from various e-commerce such as Shopee, Tokopedia, Blibli, Lazada, Lazmall, Tokopedia, Arjuna Electronics, Personal Digital, Bukalapak, Amazon, and Blanja. The median determined as the base price. After the process of cleaning and validating, 1120 rows of data left were ready to be analyzed. Table 1 denotes the number of handsets in the data. Samsung dominates the sample because it was established earlier and had published more types of smartphones than other brands. The range between the minimum and maximum price for each brand indicates the magnitude of handsets' price variation.

Table 1. Number of handsets by brand

Brands	Number of Handsets	Price in Rupiah		
		Average	Minimum	Maximum
Oppo	152	3,651,571	286,000	12,000,000
Realme	45	2,879,677	1,318,000	7,289,000
Samsung	407	4,387,389	399,000	38,000,000
Vivo	125	3,070,513	1,393,000	15,660,000
Xiaomi	391	2,872,251	454,500	11,997,000
Total	**1120**			

The selection of variables as predictors referred to previous research. Research by Ahmad et.al. indicates that brand, battery capacity, weight, operating system, RAM, memory size and display size have a significant positive effect on smartphone prices [6]. However, due to rapid technological development, some smartphone features are just appeared recently and becoming popular among users. Therefore, we added the features of the fingerprint scanner and NFC for the analysis. Table 2 summaries the predictor used in the model.

Table 2. Summaries of predictor variable

No	Variable	Description	Value
1	Brand	Brand image usually impacts the decision on purchasing a smartphone because people usually have their personal preference.	{Oppo, Realme, Samsung, Vivo, Xiaomi}
2	RAM	RAM is the memory that a phone uses in current working. Higher RAM is usually result in more responsive phone. Measured in Giga Byte (GB)	[0.278,32]
3	Internal Storage	Internal storage is used to store files such as document, pictures and videos. Large memory means more data storage capacity in the phone. (GB)	[0.158,1000]
4	Size	Screen size related to performance on the display. Larger screen enable user to watch videos, play games, browse the internet conveniently. However, phones having large screen size may be difficult to fit in the pocket. Measured in inch.	[3,7.3]
5	Density	PPI, or pixels per inch, refers both to the fixed number of pixels that a screen can display and the density of pixels within a digital image.	[133,577]
6	Main Camera	Main camera is the camera having the highest resolution on the back of the phone. Camera specification is given in MP (Mega Pixels) and higher MP indicates the ability of the camera to capture quality pictures with more fine details.	[2,108]
7	Selfie Camera	Selfie camera is basically resolution of the front camera on the phone (in MP). It enables users to make video calls and selfies.	[0.3,32]
8	Battery	The functionality of the smartphone is severely affected by battery life. Initially, the battery capacity was measured in hours (talk and standby time). It is measured in terms of Milli Amperes (mAH).	[1200,6000]
9	Weight	Design characteristics like weight (in grams) are valued by the users and can influence the demand for a phone.	[97.5,263]
10	NFC	NFC (Near Field Communication) is a short-range high frequency wireless communication technology that enables the exchange of data between devices over about a 10 cm distance. One of its popular function nowadays it for payment and top up electronic money.	{Yes, No}
11	Fingerprint	Fingerprint scanner on a phone is used for authentication and security.	{Yes, No}
12	Radio	Smartphone with radio enables the user to listen music, news, discussions on various issues, etc through FM network.	{Yes, No}

2.2. *Model*

2.2.1. Hedonic linear regression. Hedonic regression is the application of regression analysis to estimate the impact that various factors or characteristics have on the price for a good. In a hedonic regression model, the price is used as the dependent variable. Hedonic pricing first proposed by Sherwin Rosen 1974. In this study, Hedonic Price Regression implement to evaluate the impact of the smartphone's specifications on the price. The general format of the hedonic price regression is a semilogarithmic model as follow:

$$Ln(P) = \beta_0 + \beta_i C_i + e_i$$

Ln(P) is the natural logarithms of price, β_0 is the constant coefficient, β_i is the regression coefficient of characteristic variables, C_i is the characteristic variable and e_i is the error term. Estimation for parameters used the Ordinary Least Squares (OLS) method [9].

OLS method assumes the error term has zero mean [$E(e_i) = 0$], constant variance [$Var(e_i) = \sigma^2$], and zero covariance [$Cov(e_i, e_j) = 0$] [10]. A linear regression model that involves multiple explanatory variables has an additional condition to be met, i.e. the absence of multicollinearity between explanatory variables. The presence of high collinearity among the independent variables can increase the standard error of the estimated regression coefficients. The model is tested for multicollinearity using variance inflating factor (VIF).

The hedonic regression model is presented as follows:

$$\begin{aligned} Ln(PRICE_i) = \beta_0 + \sum \beta_{1i} BRAND_i + \beta_{2i} RAM_i + \beta_{3i} INTERNALSTORAGE_i + \beta_{4i} SIZE_i \\ + \beta_{5i} DENSITY_i + \beta_{6i} MAINCAM_i + \beta_{7i} SELFICAM_i + \beta_{8i} BATTERY_i \\ + \beta_{9i} WEIGHT_i + \beta_{11i} NFC_i + \beta_{12i} FINGERPRINT_i + \beta_{13i} RADIO_i \end{aligned}$$

where:
$Ln(PRICE_i)$ is natural logarithm of price of handset i
β_{ji} is regression coefficient of characteristic variable, j=1,2,…,13

2.2.2. Random forest regression. Random Forest Regression (RFR) is an ensemble learning algorithm proposed by Breiman (2001) [11]. This method constructs multiple decision trees using the different bootstrap samples of the training dataset and generalizes with the mean prediction of individual trees. The basic idea of a regression tree is examining the value of the explanatory variable and posing a binary problem that divides the node into two nodes.

In the RFR, each node is split using the best among a subset of predictors randomly chosen at that node. The randomness in choosing the predictor is to reduce the redundancy of the predictor variables and to increase the diversity of the tree in the forest. The random bootstrap sample of the training dataset for construct each tree can increase the robustness and avoid over-fitting. The final output of the RFR is concluded by averaging the prediction of each tree. RFR algorithm can describe in four steps visualized figure 2 as follows:

- Draw ntree bootstrap samples from the training dataset, where ntree is the number of trees to grow in the forest.
- For each bootstrap sample, grow the regression trees, with the following modification: choosing randomly a subset containing mtry predictors variables at each splitting mode and choose the best split based on mtry predictor variables.
- Obtain the RFR predictions on ntree decision trees regression.
- Obtain the final prediction by averaging ntree predictions in the forest.

An estimate of the error rate can be obtained based on the training dataset. For each bootstrap iteration, using the tree regression grown with the bootstrap sample to predict the data exclude the bootstrap sample (out-of-bag or OOB data). The error rate calculated by averaging the OOB predictions and it call the OOB estimate of error rate.

Several methods can be used to validate models for tuning hyperparameters in machine learning. These methods include validation, leave-one-out cross-validation (LOOCV), k-fold cross-validation (CV), repeated cross-validation. The tuning parameter method in this study uses repeated k-fold cross-validation (CV) to determine the best mtry.

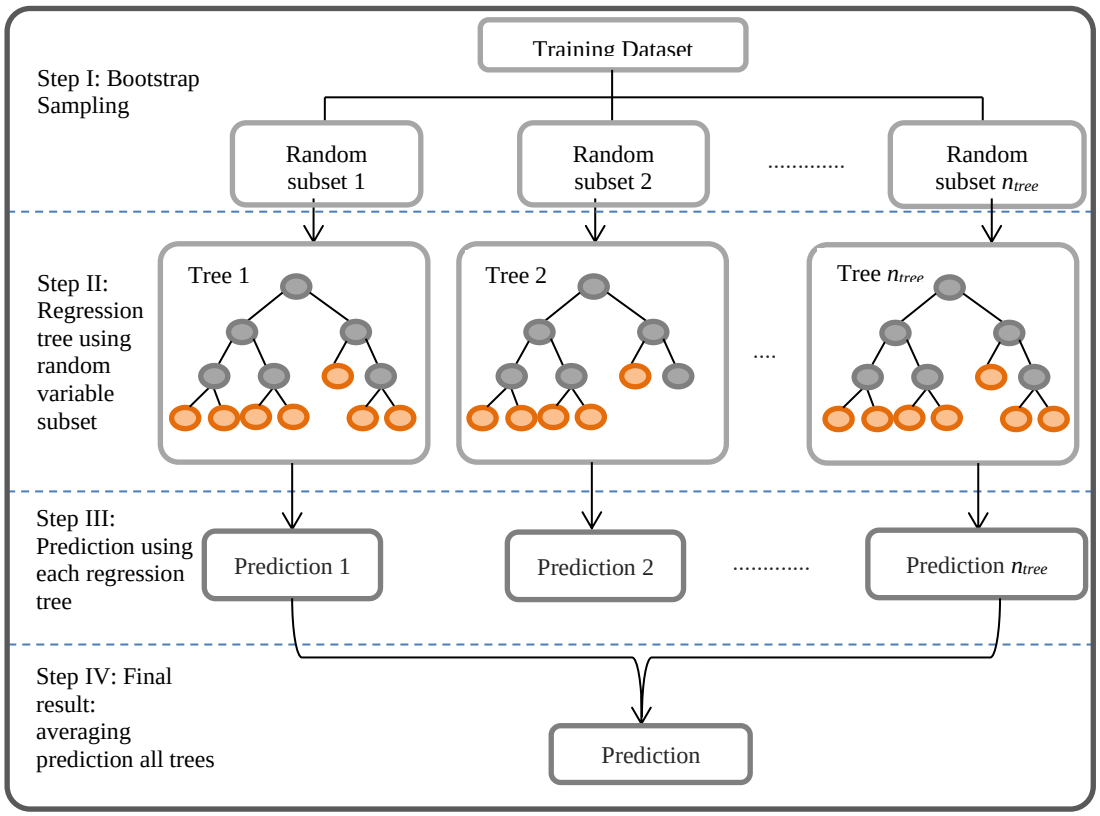

Figure2. Scheme of random forest regression algorithm

The contribution of variables in the predictive model based on variable importance (VI) metric. The permutation-based VI (PVI) is the most robust and commonly for RFR. The basic idea of the PVI metric has randomly permuted all value of variable Fi. The variable importance measure is defined as the difference in prediction accuracy caused by the permutation [11]. If a variable is not important in the prediction model, then permuted its values will not change the final model performance [12]. The commonly used predictive accuracy of the model is the mean square error (MSE). The importance of ith predictor (Fi) variable in jth tree (Tj) is defined as:

$$VI_j(F_i) = \frac{MSE\left(\hat{Y}_j, Y_j\right) - MSE\left(\hat{Y}_j^{(i)}, Y_j\right)}{MSE\left(\hat{Y}_j, Y_j\right)} \times 100\%$$

where Y_j denote OOB data of jth tree (j=1,2,…, ntree), $\hat{Y}_j$ denote the corresponding OOB predictions before permuting Fi, and $\hat{Y}_j^{(i)}$ denote the corresponding OOB predictions after permuting Fi. The total importance for variable Fi is calculated as average of VI over all trees in the forest,

$$VI(F_i) = \frac{1}{n_{tree}} \sum_{j=1}^{n_{tree}} VI_j(F_i)$$

2.2.3. Extreme gradient boosting (xgboost) regression. There are various algorithms to learn tree ensembles i.e. previously explained random forest, gradient tree boosting/ Gradient Boosting Machine (GBM), and gradient tree boosting with regularization. Extreme Gradient Boosting which often referred as XGBoost is a method of gradient tree boosting with regularization. This method initially started as a research project by Tianqi Chen in 2014, and immediately became popular due to its frequent to win machine learning competitions [13].

Boosting is an ensemble technique where new models are added to correct the errors made by existing models. Models are added sequentially until no further improvements can be made. Gradient boosting is an approach where new models are created that predict the residuals or errors of prior models and then added together to make the final prediction. It is called gradient boosting because it uses a gradient descent algorithm to minimize the loss when adding new models. This approach supports both regression and classification predictive modelling problems. Both XGBoost and GBM follows the principle of gradient boosting. There are however, the difference in modelling details. Specifically, XGBoost uses a more regularized model formalization to control over-fitting, which gives it better performance.

The name XGBoost refers to the engineering goal to push the limit of computations resources for boosted tree algorithms. The library provides a system for use in a range of computing environments, not least:

- Parallelization of tree construction using all of your CPU cores during training.
- Distributed Computing for training very large models using a cluster of machines.
- Out-of-Core Computing for very large datasets that don't fit into memory.
- Cache Optimization of data structures and algorithm to make best use of hardware [14].

Other than the computational aspects, the principles behind XGBoost follows the same idea as gradient boosting with improvements in the regularized objective. Chen et al [15] explained the XGBoost model as follows:

Assuming that a data set with n samples and m features is $D = \{(x_i, y_i): i = 1..n, x_i \in \mathbb{R}^m, y_i \in \mathbb{R}\}$. Let $\hat{y}_i$ be defined as the predict value of the model:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \qquad f_k \in F$$

where f_k represents an independent regression tree and $f_k(x_i)$ denotes the prediction score given by the k-th tree to the i-th sample. The set of functions f_k in the regression tree model can be learned by minimizing the objective function:

$$Obj = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

The l herein is a training loss function, which measures the difference between the prediction $\hat{y}$ and the object y_i. To avoid over-fitting, the term Ω penalizes the complexity of the model:

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda ||w||^2$$

Where γ and λ are the degree of regularization. T and w are the numbers of leaves and the scores on each leaf respectively. The tree ensemble model can be trained in an additive manner.

$g_i = \delta_{\hat{y}_i^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)}\right)$ and $h_i = \delta^2_{\hat{y}_i^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)}\right)$ are the first and the second order gradient on l. Let $\hat{y}_i^{(t)}$ be the prediction of the i-th instance at the t-th iteration, it needs to minimize the following objective:

$$\begin{aligned} Obj^{(t)} &= \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t(x_i)^2 \right] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \\ &= \sum_{j=1}^{T} \left[\left(\sum_{i \epsilon I_j} g_i \right) w_j + \frac{1}{2} \left(\sum_{i \epsilon I_j} h_i + \lambda \right) w_j^2 \right] + \gamma T \end{aligned}$$

where I_j denotes the instance set of leaf j. For a fixed tree structure q, the optimal weight w_j^* of leaf j and the corresponding optimal value can be calculated by:

$$w_j^* = -\frac{G_j}{H_j + \lambda}$$

$$Obj^* = -\frac{1}{2} \sum_{j=1}^{T} \frac{G_j}{H_j + \lambda} + \lambda T$$

where $G_j = \sum_{i \epsilon I_j} g_i$, $H_j = \sum_{i \epsilon I_j} h_i$, Obj presents the quality of a tree structure q. I_L and I_R are the instance sets of the left and right nodes after split. By enumerating the feasible segmentation points and selecting the minimum target function and the maximum gain partition, the gain formula is shown as follows:

$$G = \frac{1}{2}\left[\frac{\left(\sum_{i \epsilon I_L} g_i\right)^2}{\sum_{i \epsilon I_L} h_i + \lambda} + \frac{\left(\sum_{i \epsilon I_R} g_i\right)^2}{\sum_{i \epsilon I_R} h_i + \lambda} - \frac{(\sum_{i \epsilon I} g_i)^2}{\sum_{i \epsilon I} h_i + \lambda}\right] - \gamma$$

The XGBoost modelling in R utilized the "caret" package.The parameters to be tuned on this model are nrounds, max_depth, eta, gamma, colsample_bytree, min_child_weight and subsample. Nrounds controls the maximum number of iterations. Max_depth controls the depth of the tree. Eta controls the learning rate, i.e., the rate at which our model learns patterns in data. Gamma controls regularization. Colsample_bytree controls the number of features (variables) supplied to a tree. Min_child_weight refers to the minimum number of instances required in a child node. Subsample controls the number of samples (observations) supplied to a tree.

3. Result and discussion

This study used the hedonic analysis with logarithmic transformation of the price as the response. A descriptive analysis was conducted to the collected data by examining the relation of each predictor with Ln Price. Some variables indicated to be strong predictors for Ln Price than the rests. However, a variable could be a weak predictor when used alone but transform into strong predictor when combined with other predictors.

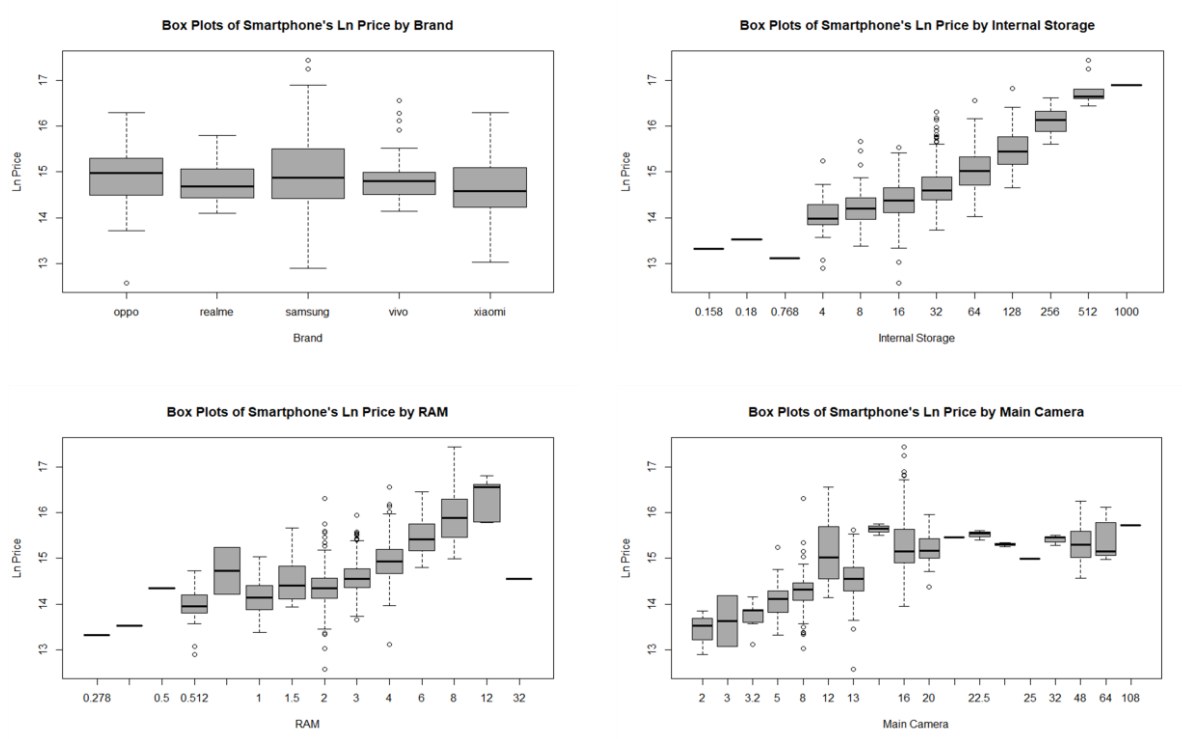

Figure 3. Distributions of smartphone's ln price data by predictor variables

Figure 3 shows box plots of the data distribution based on a five-number summary (minimum, first quartile (Q1), median, third quartile (Q3), and maximum). The first graph in Figure 3, shows that all the boxes are aligned, means that Q1, median, and Q3 on smartphone's prices are close for different brands. Price of Samsung smartphones tends to be higher than smartphones from other brands. This is indicated by the presence of a higher tail accompanied by outliers. Visually, brand alone is not a strong predictor

for the smartphone's price. The rest three graph on Figure 3 shows boxes that prices are increasing along with the increasing specifications of RAM, Internal Storage, and Main Camera so are the medians. It means that RAM, Internal Storage, and Main Camera are quite strong predictors for the price.

3.1. *Hedonic linear regression*

Hedonic regression result in Table 3 shows that all of the explanatory variables together significantly affect the price of the smartphone. It is indicated by the F test significantly rejects the null hypothesis. Adjusted R squared pointed out that 75.35 percent of price variation that can be explained by 12 explanatory variables in the model. No VIF value of each variable exceeded 10, so the conclusion is no evidence of multicollinearity in the model.

Table 3. Estimated result of ols hedonic model

Variable	Estimate Coefficient	Standard Error	t-value	VIF
Intercept	12.8010***	0.1560	82.0390	
BRAND				2.6247
Realme	-0.2555***	0.0714	-3.5790	
Samsung	-0.0604	0.0446	-1.3550	
Vivo	0.0243	0.0486	0.5020	
Xiaomi	-0.2622***	0.0411	-6.3780	
RAM	0.0411***	0.0085	4.8700	2.6726
INTERNAL STORAGE	0.0024***	0.0002	11.7080	1.8569
DENSITY	0.0016***	0.0002	8.7000	2.0171
NFC				1.9071
Yes	0.3267***	0.0338	9.6790	
SELFICAM	0.0029	0.0023	1.2960	2.1169
SIZE	0.1210***	0.0413	2.9290	5.7666
MAINCAM	0.0000	0.0013	0.0480	1.6301
WEIGHT	0.0048***	0.0010	5.1380	3.4844
BATTERY	-0.0001***	0.0000	-4.6880	3.7685
FINGERPRINT				2.1239
Yes	0.2132***	0.0373	5.7250	
RADIO				1.5337
Yes	-0.084**	0.0326	-2.5790	
R2	0.7579			
Adjusted R2	0.7535			
Ftest	171.9***			

*** and ** show statistical significance at 1 and 5 percent level of significance

The specifications of the phone camera both the front camera (selfie camera) and the back camera (main camera) have no significant effect on the price. Two of 4 the dummy variables for brand are not significantly affecting price i.e. Samsung and Vivo. It means that there is no significant price difference between smartphone from brand Samsung and Vivo with the reference brand, Oppo. In other words, the brands of Samsung, Vivo and Oppo have relatively equal prices. The negative coefficients for the dummy variables of the Realme and Xiaomi brands show that the smartphones price of these two brands is on average lower than the price of the other brands.

RAM, Internal Storage, density, NFC, size, weight, and fingerprint scanner have significantly positive effects on the price. While battery and radio have a significant effect on the negative direction

of the price of smartphones. The negative leverage of radio feature on the price could due to most recent smartphones do not have the radio FM feature because nowadays it is accessed via the internet.

3.2. Random forest regression

Two methods of machine learning regression are applied to measure variable importance: random forest regression and Extreme Gradient Boosting (XGBoost) Regression. Data processing used R software with the "caret" package by doing repeated k-fold cross-validation. The total 1120 observations were split into a training dataset (840 observations) and testing dataset (280 observations).

The process of hyperparameter tuning in random forest regression was executed using repeated k-fold cross-validation. 12 predictors of categorical and numerical type used in the model. In modelling, the categorical variables transformed into dummy variables. In the end, the total predictor variables in the model become 15. The selection of mtry parameter based on the smallest RMSE value among several subsets of variables used to separate data at each node. Figure 4 visualized the results of the repeated cross-validations process. The optimum mtry value produced is 4 with RMSE value of 0.2559.

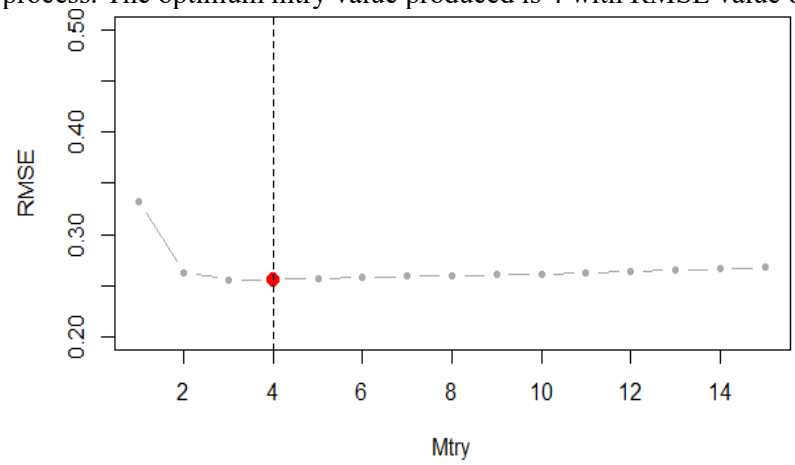

Figure 4. RMSE calculated by repeated cross-validation for tuning mtry in random forest regression

A good machine learning model is expected to do well in predicting new data (input). The performance of the random forest regression model, in general, can be seen if applying a model formed based on optimum hyperparameter on the testing dataset. Based on prediction results from testing dataset, the resulting RMSE value was 0.2509. The RMSE value in the testing dataset was not much different from the RMSE cross-validation results in the training dataset. It indicated that there was no indication of overfitting.

Figure 5 visualizes the contribution of the predictor variables to the price calculated by the permutation-based VI (PVI) metric approach. The most important specification influencing the price is RAM. It supports the initial exploration that smartphone with higher the RAM is more expensive. The second and third important variables, respectively, are Internal Storage and Density. Smartphone brand does not appear on 10 most important variables to predict price in Figure 5. It means the different brands between Oppo, Realme, Samsung, Vivo, Xiaomi, does not influence the price significantly.

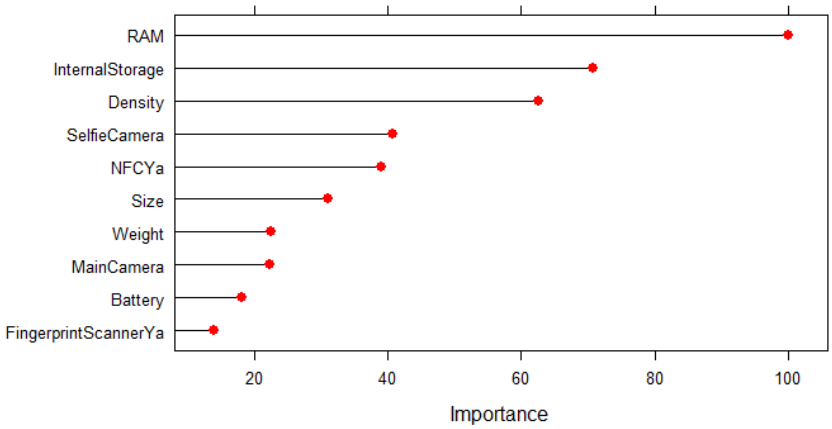

Figure 5. Variable importance based on random forest regression

3.3. Xgboost regression

The process of hyperparameter tuning for XGBoost was also conducted with repeated k-fold cross-validation using "caret" library. RMSE was used to select the optimal model using the smallest value. The final values used for the model were nrounds = 100, max_depth = 3, eta = 0.3, gamma = 0, colsample_bytree = 0.8, min_child_weight = 1 and subsample = 1. RMSE for training dataset obtained 0.2813 while the RMSE for testing dataset were 0.2536.

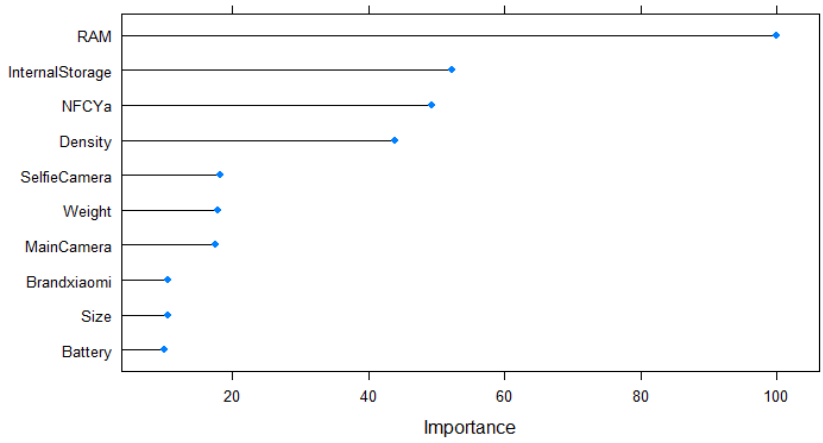

Figure 6. Variable *importance based on xgboost regression*

Figure 6 shows that the two most important variables in this model are consistent with the results from the random forest, namely RAM and internal storage. However, the third most important variables are different, NFC for XGBoost while random forest resulted in density. Smartphone brand Xiaomi does appear to influence the price on 8th rank. It is consistent with the results of the OLS regression which shows that Xiaomi brands tend to be cheaper than other brands. Specifications appeared on 10 most important predictor in both methods are RAM, Internal storage, NFC, density, selfie camera, weight main camera, size and battery. Fingerprint scanner appeared only in XGBoost, while Brand Xiaomi only appeared in random forest.

4. Conclusions

This study estimated the implicit online retail prices for 5 smartphone brands in Indonesia using the hedonic approach. Hedonic price model was used for estimating the implicit pricing by the characteristics of smartphones. RAM, Internal Storage, density, NFC, size, weight, and fingerprint scanner have significantly positive effects on the price of smartphone, while battery and radio have a

significant effect on the negative direction of the price of smartphones. Surprisingly, both main camera and selfie camera are not significant in the model. The model explained about 75 percent smartphones price variability.

Random forest and XGBoost provided consistent results that RAM is the key feature in determining smartphones price. The hedonic model also showed that not every dummy variable for brand is significantly affecting the price. This result could be input in determining the base for weighting in CPI to use RAM specifications instead of brand.

References

[1] Lancaster K J 1966 The Billion Prices Project: A new approach to consumer theory *Journal of Political Economy* 74(2), 132–157
[2] Martinez J and Garmendia 2010 Application of hedonic price modelling to consumer packaged goods using store scanner data *Journal of Business Research* 63 (2010), 690-696
[3] Gu G and Xu B 2017 Housing Market Hedonic Price Study Based on Boosting Regression Tree *Journal of Advanced Computational Intelligence and Intelligence Informatics* 21 (6) 1040-1047
[4] Deloitte Indonesia Perspectives | First Edition, September 2019. Have Indonesians' Shopping Patterns Shifted Towards Digital? [internet]. [Retrieved 2020 June 7]. Available at: https://www2.deloitte.com/id/en/pages/about-deloitte/articles/deloitte-indonesia-perspectives.html.
[5] Lin ZS and Chen CC 2013 An analysis of the economic value of smartphone in Taiwan *Journal of Asia Pacific Business Innovation & Technology Management* 003 (2013) 077-084
[6] Ahmad W, Ahmed T, and Ahmad B 2019 Pricing of smartphone attributes at the retail level in a developing country: Hedonic analysis Telecommunications Policy, Elsevier, vol. 43(4), pages 299-309
[7] Ahmad W and Anders S 2012 The value of brand and convenience attributes in highly processed food products. Canadian Journal of Agricultural Economics, vol. 60, pages 113–133.
[8] Khoirunnisa 2019 Top 5 Vendor Smartphone di Indonesia Q3-210. Retrieved on May 1, 2020, at https://selular.id/2019/11/top-5-vendor-smartphone-di-indonesia-q3-2019/
[9] Eshet T, Baron M G, Shechter M, and Ayalon, O 2007 *Measuring externalities of waste transfer stations in Israel using hedonic pricing* Waste Manag 27, 614–625
[10] Gujarati D N and Sangeetha 2007 *Basic econometrics* (4th ed.). New Delhi: Tata McGraw Hill Publishing Company Limited
[11] Breiman L 2001 *Random Forests* Mach. Learn 45 5–32, doi:10.1023/A:1010933404324.
[12] Hjerpe A 2016 Computing Random Forests Variable Importance Measures (VIM) on Mixed Continuous and Categorical Data Master's Thesis KTH Royal Institute of Technology Stockholm Sweden
[13] Chen T and He T 2014 Higgs Boson Discovery with Boosted Trees HEPML'14: Proceedings of the 2014 International Conference on High-Energy Physics and Machine Learning, Vol. 42, pages 69-80
[14] Chen T and Guestrin C 2016 XGBoost: A Scalable Tree Boosting System. *The 22nd ACM SIGKDD International Conference* DOI: 10.1145/2939672.2939785
[15] M Chen, Q Liu, S Chen, Y Liu, C Zhang and R Liu 2019 XGBoost-Based Algorithm Interpretation and Application on Post-Fault Transient Stability Status Prediction of Power System. IEEE Access, vol. 7, pp. 13149-13158. doi: 10.1109/ACCESS.2019.2893448.