Research On Laptop Price Predictive Model Based on Linear Regression, Random Forest and Xgboost

Peiru Tian

Ningbo Xiaoshi High School, Ningbo, China

ciyanke@nbufe.edu.cn

Abstract. In today's rapidly evolving technological environment, the portability and versatility of laptops have led to a significant growth in their user base as well, making them essential tools for individuals and businesses alike. Particularly since the COVID-19 pandemic, the rate of Work From Home (WFH) partially or overall WFH has exceeded 30%. Global laptop shipments have also far surpassed desktop PCs so far in 2017. Accurately predicting laptop prices is beneficial for retailers to devise competitive pricing strategies and for consumers to effectively budget and select the most suitable laptops. In this study, we used a dataset of 1320 samples to investigate the significance of features in a laptop price prediction model using linear regression, random forest, and XGBoost methods. We incorporated 13 features in the modeling process, including laptop brand, type, screen size, RAM, GPU, operating system, and weight. Three mathematical models were established to predict the price of laptop. By comparing the regression model metrics RMSE and R^2 under linear regression, random forest and XGBoost models, the RMSE under the XGBoost model is 294.11 with an R^2 of 0.85. It is evident that the XGBoost model exhibits the smallest RMSE and the highest R^2 value closest to 1. This suggests that the XGBoost model provides the highest accuracy and best fit for the predictive model.

Keywords: Random Forest; XGBoost; Linear Regression; Feature Importance.

1. Introduction

In today's rapidly evolving technological environment, the portability and versatility of laptops, have led to a significant growth in their user base as well, making them essential tools for individuals and businesses alike. From personal use to professional applications, computers have revolutionized the way we work, communicate and access information. Global laptop shipment statistics show that 161.6 million laptops were sold worldwide in 2017, surpassing desktop computers by 65.23%. Since the COVID-19 pandemic, many activities have transitioned to a work-from-home (WFH) model. According to data, in 2021, from the Central Statistics Agency (BPS) of East Java, large and medium-sized enterprises (UMB) that adopted partial remote work accounted for 32.37%, resulting in an overall remote work rate of 2.24% [1]. Laptop peripherals such as touch screen availability and user preference [2], as well as aspects like troubleshooting and efficient servicing, have also played a significant role in the widespread utilization of laptops [3].

With more and more types of laptops available in the market, the prices of different laptops also vary. There are many factors that affect the price of a laptop such as Brand, Type, Screen Size and Resolution, CPU, Laptop RAM, HD / SSD Memory, GPU, Operating System, and laptop weight. A study focusing on the Multi-Criteria Decision Making (MCDM) approach to choosing a laptop concluded that there is a significant mutual influence between laptop price and brand image, while speed, storage, display, and other/peripheral attributes have an overall influence characteristic, but not a strong one [4]. A foreign study on the brand choice behavior of college students purchasing laptop, using descriptive statistics, correlation and structural equation modeling to analyze the primary data obtained. It was concluded that the product price has a greater impact on brand choice than other considerations [5]. The prediction of the price of laptops with different configurations can be extremely helpful for both dealers and customers. It not only helps dealers to price their products better, but also helps consumers to calculate their budgets efficiently. In order to be able to determine the reasonable price of laptops, various prediction models have been proposed in the literature, such as classification methods based on decision tree algorithms, which utilize machine learning tools to

predict the sales of laptops caused by various types of operating system factors in real applications [6].

This study is based on a sample data of 1320 laptop prices with different configurations. Firstly, the data were preprocessed, involving the removal of missing values and the elimination of factors that held no significance for the analysis, thereby streamlining the dataset. Secondly, the prices of laptops were predicted by building Random Forest, XGBoost and Linear Regression models. Concurrently, the RMSE and R2 regression model metrics were used to assess the predictive values, aiding in the identification of the most accurate model. Compared with other research efforts, this study has a more comprehensive dataset, which contains a variety of factors. Moreover, it establishes a variety of mathematical models and tests and filters them. The predictive value for the price of laptop ultimately obtained in this study will be more accurate.

2. Method

2.1. Dataset

In this paper, a dataset comprising 1320 samples is employed to predict the prices of laptop computers. The dataset originates from the Kaggle website and is specifically named "Laptop Price", containing 13 feature variables that cover various aspects of laptops in different configurations. These characteristics include brand, laptop type, RAM, screen size and resolution, CPU, hard disk/solid state drive storage, graphics processing unit (GPU), operating system, and weight. By utilizing this dataset, the research is capable of predicting the prices of laptops with different configurations. Additionally, the study establishes multiple mathematical models for analysis and comparison. The large size of this dataset, which contains a rich set of characteristic variables, allowed the study to analyze and predict laptop prices in a more comprehensive manner.

2.2. Preiction Algorithm

In order to enhance price prediction accuracy, the Random Forest, XGBoost, and Linear Regression models were selected to forecast laptop prices in this paper [7]. The performance of these three models is compared using metrics such as RMSE and R^2.

2.2.1 Random Forest

Random Forest is a powerful machine learning algorithm for solving both classification and regression problems. The core idea is to integrate the predictions of multiple decision trees in order to reduce the risk of overfitting and obtain more stable predictive performance. Each decision tree is trained on different subsamples and subset of features to increase the diversity of the model. The basic workflow of the Random Forest algorithm includes steps such as data preparation, random sampling, feature selection, decision tree construction, integrated prediction, and feature importance assessment, and model evaluation.

2.2.2 XGBoost

XGBoost is a machine learning algorithm based on gradient boosting trees, which constructs powerful predictive models by integrating multiple decision tree models. XGBoost offers the advantages of high performance, scalability, and flexibility, as well as features such as feature importance evaluation and parallel computing, which enable it to handle large-scale datasets and prevent overfitting.

Typically, XGBoost is used to solve problems such as classification, regression, and ranking. It achieves this by iteratively training multiple weak learners, namely decision trees. Each iteration corrects the model's error based on the previous iteration's prediction results [8]. The XGBoost objective function is comprised of a loss function and a regularization term, with the loss function represented by Equation 1.

$$Loss\ Function = \sum(y_i - \hat{y}_i)^2 \quad (1)$$

Where y_i is the true label, and $\hat{y}_i$ is the predicted value of the model.

In order to control the complexity of the model and avoid overfitting, XGBoost introduces regularization terms, including L1 regularization and L2 regularization. The regularization terms are represented by Equation 2.

$$Regularization\ term = \lambda \times \sum |w_j| + 0.5 \times \lambda \times \sum {w_j}^2 \quad (2)$$

where w_j is the weight of the model and λ is the regularization parameter used to balance the effects of the loss function and the regularization term.

2.2.3 Linear Regression

Establishing a linear relationship between an independent variable and a dependent variable is commonly done using the statistical technique of linear regression. It attempts to fit the data by reducing the difference between the projected and actual observed values under the presumption that the independent and dependent variables are correlated linearly. The least squares approach is used in the linear regression model to estimate the independent variable coefficients. By minimizing the sum of squared residuals between the observed values and the predicted values of the model, the least squares approach seeks to identify the best fit line or hyperplane. Using the estimated coefficients, it becomes possible to predict new values of the dependent variable and explain the degree of influence of the independent variable on the dependent variable. The equation can be represented as Equation 3.

$$y = b_0 + b_1 \times x \quad (3)$$

If b0 denotes the intercept and b1 denotes the slope, and x and y are the independent and dependent variables. The linear relationship between the independent variable x and the dependent variable y is described by this equation. To produce predictions and inferences, the best intercept and slope values can be estimated by fitting the data.

2.3. Metrics

In this research, two typical metrics called RMSE and R^2 are employed. The Root Mean Square Error (RMSE) is a commonly used metric for assessing the accuracy of predictive models, often used in the fields of statistics and machine learning. It is suitable for evaluating regression models that aim to predict continuous numerical values. RMSE offers a measurement of the average deviation between the predicted and actual values and quantifies the difference between the model's projected values and the actual observed values. By computing the square root of the mean of the squared differences between the predicted values and the actual values, RMSE evaluates the precision of a predictive model. The accuracy of the model is indicated by a decreased RMSE number, with the predicted values of the model being nearer to the actual values. Equation can be utilized to calculate RMSE.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} ||y_i - \hat{y}_i||^2}{N}} \quad (4)$$

R^2 is a statistical metric that assesses how well a regression model fits the data. It represents the percentage of the dependent variable's variance that the independent variable in the regression model can account for. It is a number between 0 and 1, with a closer value to 1 indicating a better match and a closer value to 0 indicating a worse fit.

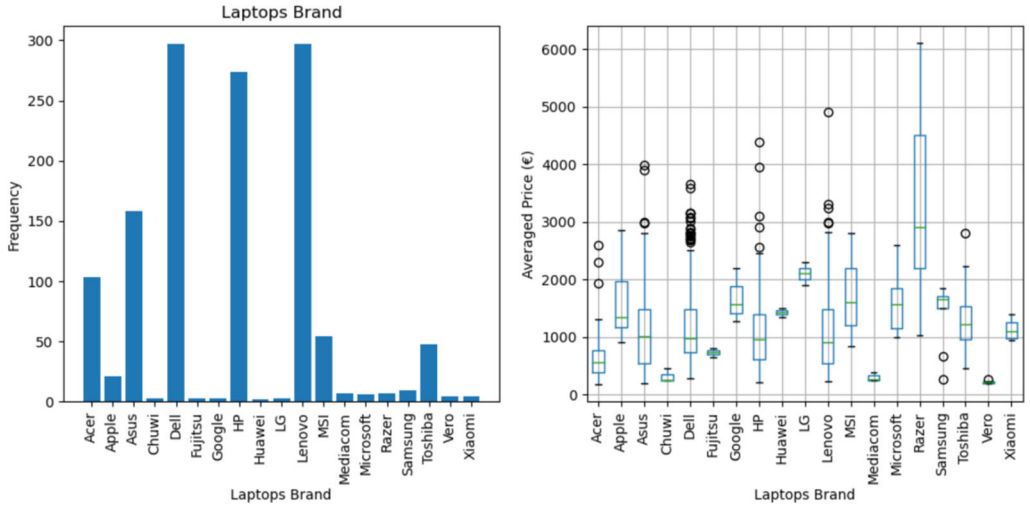
3. Result and discussion

3.1. Brand Analysis

The data were grouped according to the brands of laptop to compare the market purchase frequency and average price of different brands. In Figure 1(a), the x-axis represents various laptop brands,

while the y-axis denotes purchase frequency. The bar chart in the figure shows the purchase frequency of each brand and helps us to analyze the popularity of each brand in the market. From Figure 1(a), it is evident that DELL, Lenovo, and HP exhibit significantly higher sales than other brands, indicating that laptops from these brands hold a substantial share in terms of purchase frequency. Conversely, brands like Chuwi, Fujitsu, Google, Huawei, and LG, represented by shorter bars, indicate lower purchase frequency and likely have a relatively minor impact in the market.

This paper also plots box plots based on laptop brands and their purchase prices using seaborn, as shown in Figure 1(b). The x-axis represents different computer brands, while the y-axis represents the transaction prices of each laptop. Based on the length of the whisker line, we can determine the degree of data dispersion. If the whisker line is longer, it means that the data are more discrete and there is a large range of data. While the whisker line is shorter, it means that the data are more concentrated. As can be seen from Fig. 1(b), Asus, Dell, HP, Lenovo, and Razer have longer whisker lines, which indicates that there is a wide range of transaction prices for laptops of this brand. In addition, by examining the medians of the box plots, we can observe that not all medians are close to the middle of the box. This suggests that the distribution of transaction prices for certain brands is relatively uneven.

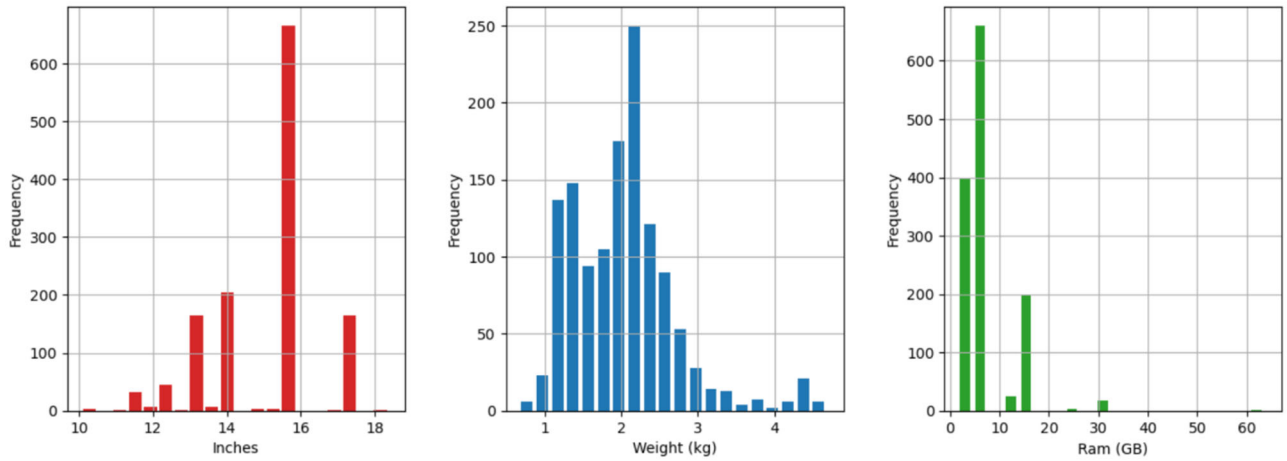


(a) (b)

Fig. 1 Brand Analysis

3.2. Analysis of Laptop Features

The relationship between the laptop purchase frequency in the market and the laptops size, weight and RAM has been analyzed in Figure 2 in terms of data statistics and bar chart analyses, shedding light on consumer preferences when buying laptops. We can observe from Figure 2(a) that 15-inch computers are the most preferred by consumers, with 13-inch, 14-inch, and 17-inch also holding some market share. From Figure 2(b), we can see that the weight of laptop is mainly concentrated in 1-3kg, with 2.2 kg being the most favored weight. This indicates that portability plays a crucial role in consumers' purchasing decisions when choosing laptops. Laptops weighing over 3.5 kg may be used as gaming notebooks or mobile workstations with more powerful performance and hardware configurations, and are designed for professional users, so weight is not a major consideration. As depicted in Figure 2(c), RAM is mainly concentrated in 8G, 4G, and 16G capacity sizes. For consumers who need to run multiple applications at the same time or handle substantial tasks, they will also choose 32G RAM capacity according to their demands.

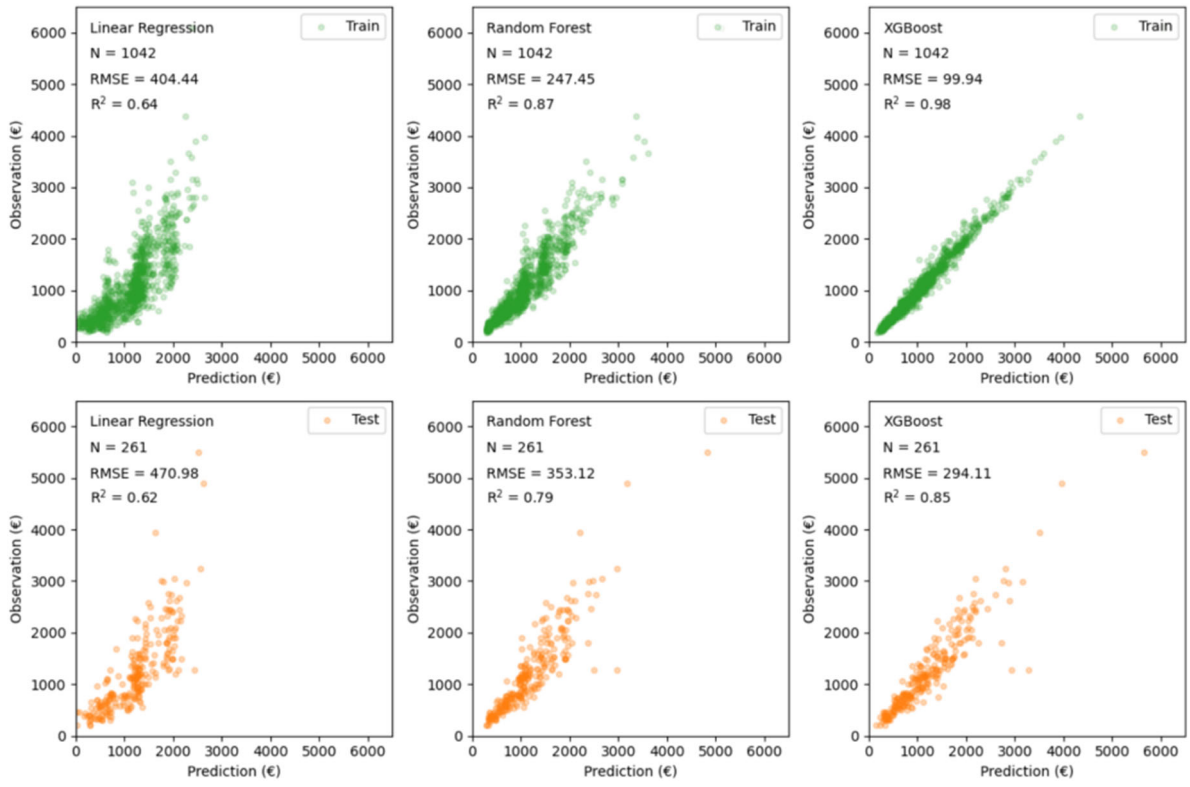


(a) (b) (c)

Fig. 2 Data Histogram According to Different Features

3.3. Model Performance

The data scatter plots corresponding to the linear regression, random forest and XGBoost models are given in Fig. 3. The original sample dataset was divided into a training set of 1042 data points and a test set of 261 data points to ensure the diversity and reliability of the dataset. Under the linear regression model, the RMSE is 470.98 with an R^2 of 0.62, as shown in Fig. 3 (a). Under the random forest model, the RMSE is 353.12 with an R^2 of 0.79, as depicted in Fig. 3 (b). Under the XGBoost model, the RMSE is 294.11 with an R^2 of 0.85, as presented in Fig. 3 (c). Comparing the regression model metrics RMSE and R^2 under linear regression, random forest and XGBoost models, it is evident that the XGBoost model exhibits the smallest RMSE and the highest R^2 value closest to 1. This suggests that the XGBoost model provides the highest accuracy and best fit for the predictive model [9].

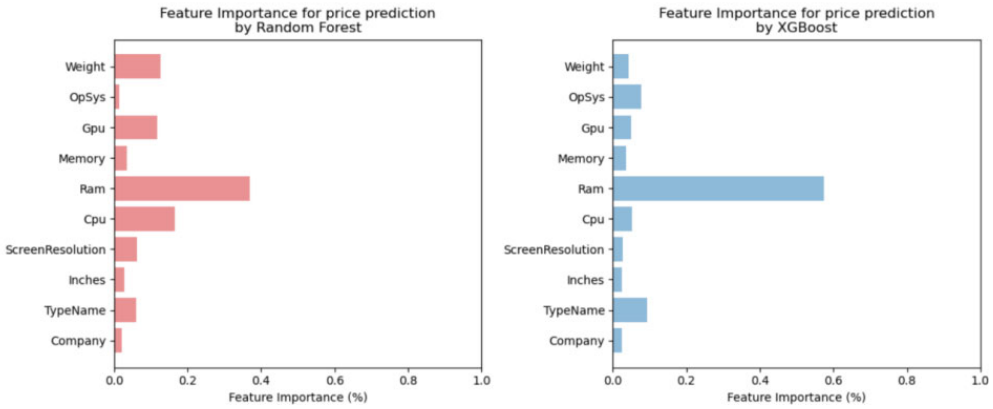


(a) (b) (c)

Fig. 3 Prediction Performance of Different Performance

3.4. Feature importance

The dataset used in this thesis contains 13 feature variables. The Random Forest and XGBoost were used to analyze the relevant laptop data to determine which features such as laptop weight, operating system, GPU, Memory, RAM, and CPU are most crucial for predicting laptop prices. Feature importance is a metric employed to gauge the significance of each feature in predicting the target variable. Ten feature variables are selected in Figure 4 and calculated a statistical value for the importance of each feature, which is used to measure the extent to which the feature contributes to the model's predictions. It can be concluded from Figure 4(a) that when using the Random Forest model for feature selection, RAM, CPU, weight, and GPU hold relatively high importance ratios. The results using the XGBoost model is given in Figure 4(b), with RAM's feature importance significantly outweighing other features. Both the Random Forest and XGBoost models show a high feature importance ratio for RAM, indicating its substantial influence on the laptop price prediction model's outcomes.



(a) (b)

Fig. 4 Feature Importance Ranking by Random Forest and XGBoost

4. Conclusion

With the popularity of laptops, the prices of laptops with different configurations vary. This study investigated laptop prices across a range of configurations, accounting for elements like laptop brand, type (Gaming, Notebook, Ultrabook, etc.), the graphics processing unit (GPU), screen size and resolution, hard disk/SSD Memory, the central processing unit (CPU) laptop RAM, operating system and weight. Multiple mathematical models were established to predict laptop prices, including linear regression, random forest, and XGBoost models. Through RMSE and R2 tests and comparisons, the XGBoost model has the smallest RMSE value and the R2 value closest to 1, indicating the highest accuracy and suitability for this prediction. Furthermore, both dealers and consumers are greatly helped by this study. For dealers, predicting laptop prices enables them to develop competitive pricing strategies. By analyzing market trends and predicting future prices, they can adjust their pricing to attract customers and maintain a competitive edge. For consumers, predicting laptop prices can assist them in planning their budget effectively and comparing different models to choose the laptop that best fits their budget.

However, this study still has certain limitations, such as insufficient analysis of the demand for laptop prices among different consumer groups and insufficient analysis of the impact of time factors on laptop prices, etc. In the subsequent study, predictive models will be based on more comprehensive datasets and more comprehensive feature variables [10]. Comparing the prediction performance of multiple models, we will obtain more reasonable and accurate prediction models.

References

[1] Siburian Astri Dahlia, Sitompul Daniel Ryan Hamonangan, Sinurat Stiven Hamonangan, et al. Laptop Price Prediction with Machine Learning Using Regression Algorithm. Jurnal Sistem Informasi dan Ilmu Komputer Prima (JUSIKOM PRIMA), 2022, 6(1): 87-91.

[2] Ou, Yang-Kun. User preference and usability assessments of touchpad surface tactile in laptops. Human Factors and Ergonomics in Manufacturing & Service Industries, 2020, 30(4): 311-317.

[3] Mahapatra, Sabita, Saumya Sharma. My experience of laptop purchase. Emerald Emerging Markets Case Studies, 2016, 6(2): 1-13.

[4] Sönmez Çakır, Fatma, and Mehmet Pekkaya. Determination of interaction between criteria and the criteria priorities in laptop selection problem. International Journal of Fuzzy Systems, 2020, 22(4): 1177-1190.

[5] Rai Bharat, Budhathoki Prem Bahadur. Factors affecting brand choice behavior of laptop purchases of university students in Nepal. Cogent Arts & Humanities, 2023, 10(1): 2194126.

[6] Gulzat Turken, Lyazat Naizabayeva, Siládi Vladimir, et al. RESEARCH ON PREDICTIVE MODEL BASED ON CLASSIFICATION WITH PARAMETERS OF OPTIMIZATION. Neural Network World, 2020 (5): 295-308.

[7] Kushwaha Aaryan, Bansal Vasu, Goti Abuzar Shaikh, et al. Optimised Laptop Price Prediction. Journal of Computer Science Engineering and Software Testing, 2023, 9(1): 25-31.

[8] Dong Jianwei, Chen Yumin, Yao Bingyu, et al. A neural network boosting regression model based on XGBoost. Applied Soft Computing, 2022, 125: 109067.

[9] Zamani Joharestani M, Cao C, Ni X, et al. PM2. 5 prediction based on random forest, XGBoost, and deep learning using multisource remote sensing data. Atmosphere, 2019, 10(7): 373.

[10] Zafar Jean-Denis, Stéphanie Himpens. Webscraping Laptop Prices to Estimate Hedonic Models and Extensions to Other Predictive Methods. 16th meeting of the Ottawa Group on Price Indices, Rio de Janeiro. 2019.