Proceedings of the 2nd International Conference on Financial Technology and Business Analysis
DOI: 10.54254/2754-1169/46/20230348

Mobile Phone Price Range Prediction Based on Machine Learning Algorithms

Jun Wang[1,a,*]

[1]Big Data Management and Application, Beijing University of Posts and Telecommunications, BeiTaiPingZhuang, Beijing, China
a. wj2021@bupt.edu.cn
**corresponding author*

Abstract: Mobile phones with different price points on the market have different parameter configurations, which provide a lot of information related to the price. Mobile phone prices are also the focus of attention of the mobile phone market, manufacturers and consumers. This study used machine learning technology to investigate the issues in mobile phone sales, mainly the prediction of mobile phone price range and the feature contribution of the price range. The data resource includes the release price of mobile phones in the European market from 2018 to 2021. The data augmentation technologies are also applied in this study. Among the three established machine learning models, the accuracy rate of the training set of the Catboost model exceeds 80%, which is the model with the highest accuracy rate. The accuracy rate of Bagging Classifier is slightly lower than that of CatBoost, but the prediction result is the most stable. Such a result conforms to the characteristics of the algorithm itself. The conclusion of the analysis of feature contribution of numerical data to accuracy rate is that rear camera pixels, number of cameras and screen size have negligible impact on model prediction accuracy, which means feature contribution is minimal. The research results are reasonable and can provide references for merchants to set prices and for users to purchase mobile phones.

Keywords: mobile phone price prediction, feature analysis, bagging classifier, CatBoost, naive bayes

1. Introduction

The proliferation of the economy and advancements in technology have rendered mobile phones have become an indispensable tool in the daily life, work, study, and entertainment activities of the majority of individuals. The mobile phone market also presents fierce competition and diversified development trends with the competition of old-brand manufacturers and the entry of different emerging manufacturers. Mobile phones of different brands, models, and performances have emerged. The price of mobile phones is subject to the influence of these multifarious factors. Varied prices have a consequential impact on consumers' purchasing decisions and reflect manufacturers' market strategies. Consequently, the ability to predict mobile phone prices reasonably and accurately holds significant implications for both consumers and manufacturers. Consumers can analyze cost-effectiveness based on price predictions and choose the right mobile phone; manufacturers can formulate reasonable pricing strategies based on price predictions to improve market competitiveness and profit margins.

Historically, many studies used to employ economic theories and economic phenomena to investigate the influence of factors such as mobile phone attributes and prices on consumers' willingness to purchase through investigations on consumer purchase behavior or to study the impact of factors such as competition among manufacturers on mobile phone technology iterations [1]. However, such analysis and prediction methods are affected by the uncertainty of the market environment and the subjective factors of the forecaster, which may lead to instability and inaccuracy of the forecast results. Currently, computer technology is rapidly developing, and machine learning technology has been widely applied in fields such as image processing, natural language processing, and artificial intelligence.

Machine learning algorithms can perform comprehensive comparisons across multiple features in a fraction of the time, and they are applied for complex problems in many fields such as finance, education, industry, medicine and e-commerce. The algorithms are commonly used methods to predict the price of new items through the feature attributes of items. Mihir et al. once predicted the price of diamond specimens through attribute values. They used the Catboost regression algorithm to build a regression model with the help of attribute values extracted from images for prediction [2]. To predict mobile phone class price, Asim et al. used Naive Bayes Classifier to predict and analyzed the contribution of each attribute [3]. Kofi Nti et al. employed the Bagging algorithm and compared it with other classifiers to evaluate the performance of the Bagging Classifier when evaluating ensemble learning for stock market prediction [4]. The selection of appropriate predictive models is contingent upon the specific research topics and dataset characteristics. Many studies also apply different machine learning models to predict mobile phone prices. Different predictive models are suitable for different research topics and data. Nasser et al. built an artificial neural network (ANN) model to predict the price range of mobile phones using a dataset containing information about mobile phones with 96.31% accuracy [5]. Çetın et al. used different classification algorithms such as Random Forest Classifier, Logistic Regression Classifier, Decision Tree Classifier, and other methods to predict mobile phone price class and compared the performance of each model [6].

However, certain gaps exist in the current body of literature pertaining to the prediction of mobile phone prices. Specifically, some studies have neglected to examine the individual contributions of predictors, while others have failed to explore the distinctive characteristics of multiple models. This study will use feature engineering methods in machine learning to analyze mobile phone parameter configurations and predict mobile phone price ranges and the feature contribution of price ranges by establishing a model. This study uses the following three different machine learning prediction methodologies, namely Bagging Classifier, Catboost, Naive Bayes. Finally, this study compared the prediction accuracy and prediction effect of the three models, and analyzed the optimal model group Bagging in predicting the price of mobile phones in this set of data sets. This study also analyzed the contribution of each low feature to the prediction accuracy of each model which can help select the appropriate forecasting model and predictors.

2. Method

2.1. Dataset Description and Preprocessing

This study used the dataset provided by a dataset on Kaggle [7]. The dataset encompassed a total of 407 data entries, encompassing comprehensive information regarding the specifications and initial launch prices of multiple mobile phone models from diverse brands, within the European market, throughout the temporal scope of 2018 to 2021. The 8 features includes details for each device: Brand, Models, Storage, RAM,Screen Size (inches),Camera (MP),Battery Capacity (mAh), and Price ($).

2.1.1. Data Cleaning

The process of data cleaning contains three parts. First, removing units such as "GB" from the Storage and RAM variables, as well as the elimination of "MP" from the Camera (MP) attribute. This facilitates for subsequent analysis of the value. Subsequently, adding a new variable at the same time: the number of cameras (n_cameras), this is the value in the variable "camera", that is, the accumulated pixels of each camera is split to calculate the number of cameras equipped on each device. Second, discarding the explanatory content such as the unit of each variable, and rename them. All variable names after renaming are: brand, model, storage, ram, screen_size, battery, price, n_cameras. Finally, converting the feature to a numeric type and remove missing values as well as duplicate data. There are still 406 items of data in the data frame. Figure 1 presents the price density before and after removing duplicate values.

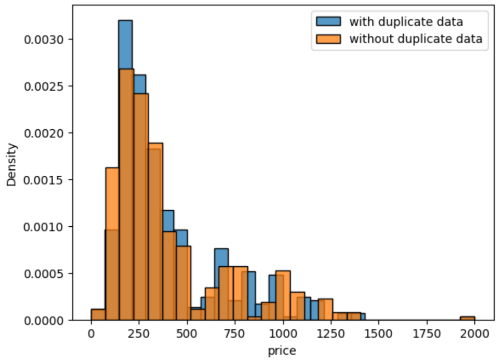

Figure 1: Price Density of the collected dataset (Photo/Picture credit: Original).

2.1.2. Correlation Analysis

This stage divided the mobile phone prices into four ranges according to quartiles and analyzed the correlation between each feature and the price range. The quartiles of the calculated prices are 199, 289, 469, 1999, and the price range is set as 0: (0-199), 1: (200-289), 2: (290-469), 3: (470-1999). Then the degree of correlation between the numerical variable and the price range is calculated. Table 1 is the results of the correlation analysis.

Table 1: Price density of the collected dataset.

Feature Name	Correlation Degree
ram	0.64
storage	0.62
res2	0.51
res3	0.47
res1	0.28
res4	0.17
n_cameras	0.05

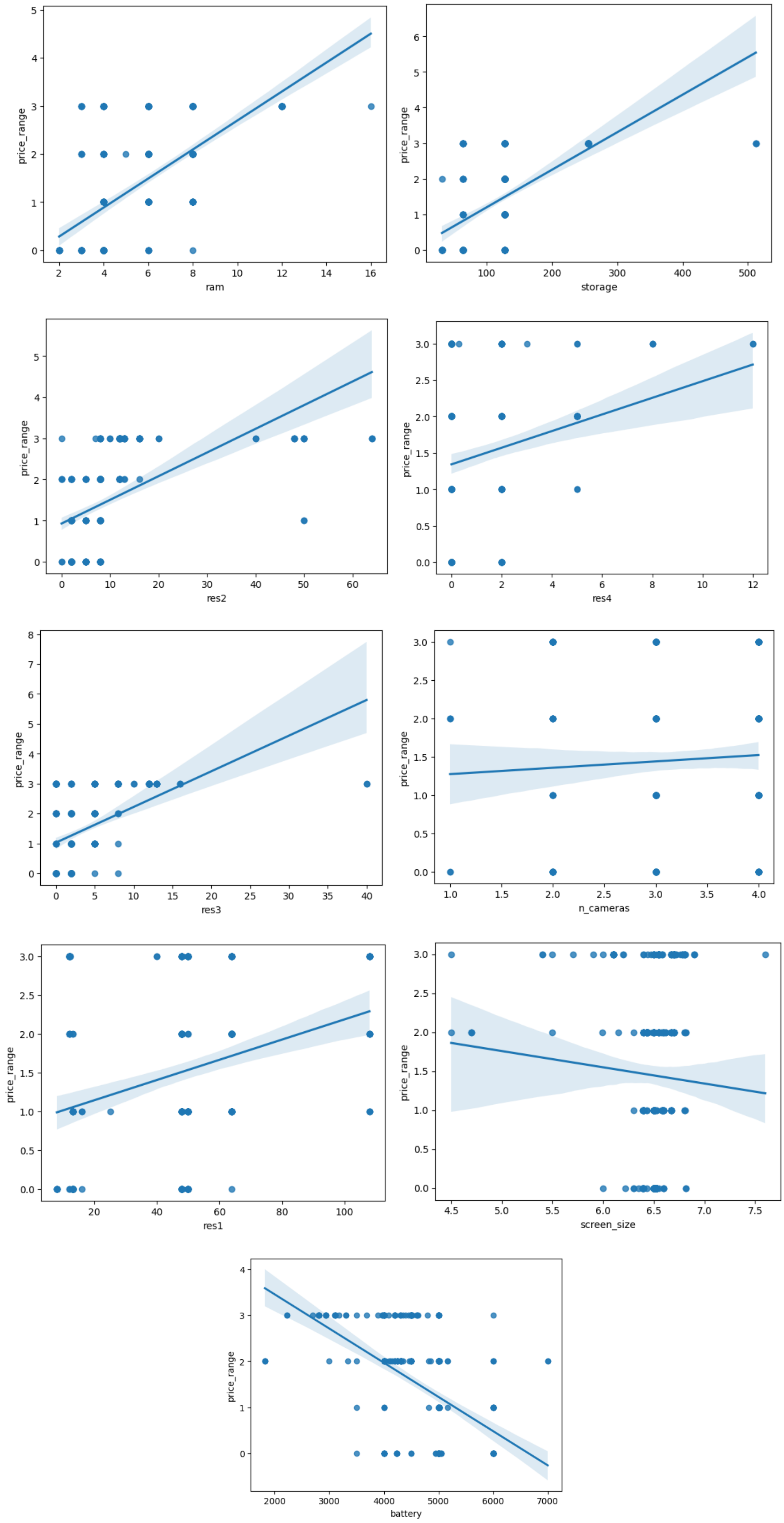

Figure 2: Correlation analysis of each feature and price_range (photo/picture credit: original).

According to the correlation analysis and the figures shown in Figure 2, the two indicators res4 and n-cameras have a low correlation with the price range, while the screen size and battery capacity correlate negatively.

2.2. Machine Learning Based Models

The main steps of this section are shown in Figure 3, which describes the process of modeling predictive analysis.

Build Model
Fit the Model
Performance Evaluation

Figure 3: The framework of the study (photo/picture credit: original).

First, the research step requires modeling. Three models of Bagging, Catboost, and Naive Bayes are selected here. Second, the model is fitted using the training set, which is 90% of the total data partition selection. Third, use the test set to evaluate the prediction ability of the model, and the test set is 10% of the total data partition selection. The evaluation indicators include accuracy, RMLSE, MSE, and MAE.

2.2.1. Bagging Classifier

Bagging Classifier is an ensemble meta-estimator that improves the accuracy and stability of the model by combining the predictions of multiple base classifiers. It fits base classifiers on each random subset of the original dataset [8]. In this study code, the base classifier is a decision tree classifier, and each base classifier is trained using 90% of samples and 90% of features. The final prediction result is obtained by voting or averaging the predictions of all base classifiers. This is the result of fitting the Bagging Classifier model while using all attributes to predict the price_range.

2.2.2. CatBoost

The CatBoost algorithm is the third improved algorithm based on the Gradient Boost Decision Tree (GBDT) algorithm after XGBoost and LightGBM, and it has advantages in processing classification features and predicting offsets. CatBoost has high accuracy and the ability to handle categorical features. It minimizes the loss function by iteratively adding decision trees, each fitting a new decision tree on the residuals from the previous iteration [9]. It can effectively reduce model overfitting and improve prediction performance. Since the mobile phone attributes include discrete and disordered battery capacity, screen size, and other categorical features, using the CatBoost algorithm may learn more information to a greater extent, thereby improving model performance. This is the result of fitting the CatBoost Classifier model while using all attributes to predict the price_range.

2.2.3. Naive Bayes

The Naive Bayes algorithm is a classifier model based on Bayesian decision theory, which can use known categories of data to train the model to achieve category judgment for unknown categories of data [10]. It can use the information in the data set to predict the price of the mobile phone according to the configuration attributes of the mobile phone, but it assumes that the features are independent of each other. It is expressed in the formula as

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{1}$$

2.2.4. Evaluation Metrics

The study uses accuracy to measure how well the classification model can correctly predict the class labels of a set of samples. The formula of accuracy is:

$$\text{accuracy} = \frac{\text{number of correct predictions}}{\text{total number of predictions}} \tag{2}$$

Next, the study used Root Mean Square Log Error (RMSLE), Mean Square Error (MSE) and Mean Absolute Error (MAE) to evaluate the performance of the regression model, which measures the difference between the predicted value and the true value of the target variable. The smaller the values of these three indicators, the better the performance of the regression model.

RMSLE is useful when more penalties are imposed on underestimation or when the target variable has a wide range of values. The calculation formula is as follows:

$$\text{RMSLE} = \sqrt{\text{mean}((\log\left(y_{pred} + 1\right) - \log\,(y_{true}} + 1))^2 \tag{3}$$

where y_{pred} is a vector of predicted values, y_{true} is a vector of true values.

MSE is used to penalize large errors more.The calculation formula is as follows:

$$\text{MSE} = \text{mean}((y_{pred} - y_{true})\,^2) \tag{4}$$

MAE is used to treat all errors equally regardless of their magnitude. The calculation formula is as follows:

$$\text{MAE} = \text{mean}(|y_{pred} - y_{true}|) \tag{5}$$

2.2.5. Contribution of Features

The features of the mobile phone configuration in the data set are multivariate, so when training the machine learning model, some features with a correlation of less than 0.3 may appear. These features may have a small impact on prediction accuracy and may be unimportant. This part will select the features whose correlation with the price range is less than 0.3 and analyze the impact of these features on the accuracy of the model by adding these features to the prediction model one by one or deleting these features from the model one by one [3]. When adding or removing features with a small degree of correlation one by one, proceed from large to small according to the degree of correlation, namely res1 (0.18), res4 (0.17), n_cameras (0.05), screen_size (-0.05). When removing features one by one, delete them from all predictors used, and when adding features one by one, start with all the features with high correlation, that is, ram (0.64), storage (0.62), res3 (0.47), battery (-0.5).

3. Results and Discussion

Table 2 demonstrated that CatBoost achieved the highest prediction accuracy. Bagging is second but only three to five percentage points behind CB, and the three values of RMSLE, MSE, and MAE are the lowest, which means that the Bagging Classfier model is the most stable. This assertion is further supported by Figure 4, which depicts the fitting effects of the models. It is observed that the Bagging classifier exhibits the least number of scattered points deviating from the fitted line. This phenomenon can be attributed to CatBoost's ability to reduce the model’s bias, resulting in a high accuracy rate. Moreover, Bagging Classifier and CatBoost may be because it can effectively reduce the variance of the model, so it has high stability. Naive Bayes perform poorly since there are many feature items with high correlation among the ten selected mobile phone configuration parameters, and the Naive Bayes model assumes that the features are independent of each other and does not consider the linear and nonlinear relationships between features which can lead to model inaccuracy and instability [11]. Moreover, compared with the CatBoost and Bagging models, which can better utilize feature information through adaptive learning and parameter adjustment, the Naive Bayes model does not have a parameter adjustment method for adaptive learning, and insufficient or too high parameters lead to a decrease in model accuracy.

Table 2: Performance of diverse models evaluated by various metrics.

Models	Train				Test			
	Accuracy	RMSLE	MSE	MAE	Accuracy	RMSLE	MSE	MAE
Bagging Classifer	0.77112	0.14664	0.07836	0.06583	0.74999	0.17026	0.08333	0.08333
CatBoost	0.85714	0.13395	0.08333	0.08333	0.80873	0.31274	0.40625	0.21875
Naive Bayes	0.71771	0.29793	0.40439	0.31661	0.72143	0.36950	0.75000	0.41667

Table 3: Contribution of features when removing features.

Features	Accuracy of Models			
	Relevance	CatBoost	Bagging Classifier	Naive Bayes
No features removed	-	0.80873	0.80262	0.70243
res1	0.28	0.74301	0.74301	0.66136
res4	0.17	0.73681	0.75551	0.67078
n_cameras	0.05	0.74931	0.74925	0.65222
screen_size	-0.05	0.74618	0.74949	0.66159

Table 4: Contribution of features when adding features.

Features	Accuracy of Models			
	Relevance	CatBoost	Bagging Classifier	Naive Bayes
No features added	-	0.73050	0.68973	0.63953
res1	0.28	0.77425	0.78671	0.71166
res4	0.17	0.77425	0.80868	0.68333
n_cameras	0.05	0.78688	0.79613	0.70537
screen_size	-0.05	0.79305	0.76478	0.70218

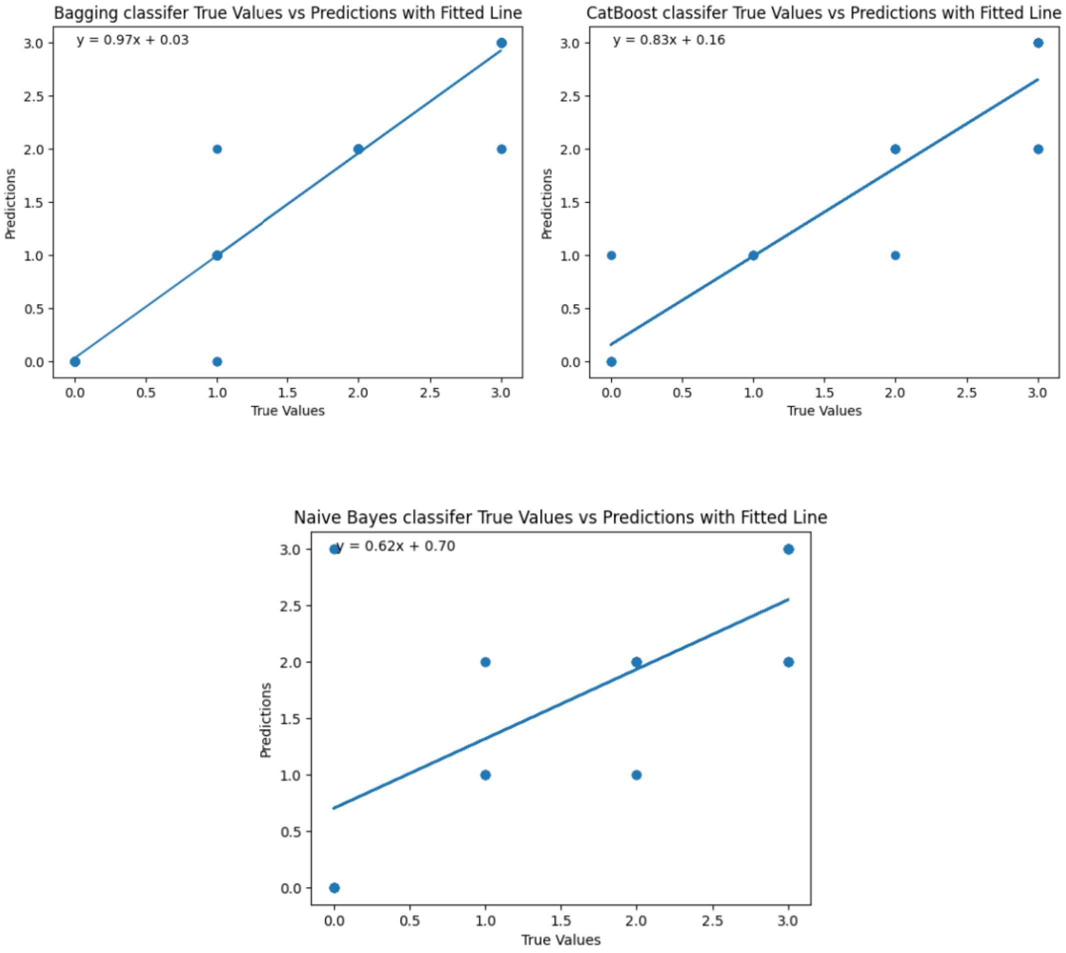

Figure 4: Fitted line based on different models (photo/picture credit: original).

The findings presented in Table 3 indicate that the removal of the feature res1, which exhibits the highest correlation coefficient (0.28) among the four features considered, led to a significant decrease in the accuracy rates of all three models. After removing the small res4 (0.17), n_cameras (0.05), and the negatively correlated feature screen_size (-0.05), the accuracy of the model may increase or decrease each time it runs, but they are all within 1%. It is due to the randomness of the data, suggesting that these three features exert minimal influence on mobile phone price prediction.

Table 4 shows the changes in model prediction accuracy when adding these four features individually. Among them, res1 emerges as the most influential. After adding, the accuracy of the three models has increased by 4%-9%, and it has the most considerable correlation with it (0.28) related. However, after adding the three features of res4, n_camera, and screen_size, the accuracy of the three models increases or decreases. However, the changes are irregular within 2.3%, possibly due to the small amount of data and random fitting data. However, these irregular changes further support the notion that these three features exert a relatively minor influence on the prediction of mobile phone prices.

4. Conclusion

This study employs three machine learning algorithms, namely Bagging Classifier, CatBoost, and Naive Bayes, to predict mobile phone price ranges and analyze the contribution of mobile phone configuration parameters. The findings of this research indicate that within the scope of this investigation, the Bagging Classifier exhibits a high prediction accuracy ranging from 78% to 80%. It demonstrates good stability and outperforms both CatBoost and Naive Bayes models. The dataset's three features (i.e. rear camera pixels, screen size, and the number of cameras), whose correlation with the price range is less than 0.2, have less impact on the predictive model. Since the data set used in this paper is small, and the research focuses on the comparison of multiple models without focusing on improving the prediction effect of each model, the effect of training the model may need to be better. Nevertheless, the process adopted in this study, including the modeling algorithm used, can be applied to similar price forecasts in other fields, such as food, automobiles, housing, etc., can aid producers, sellers, and consumers in making informed decisions regarding pricing and purchases.

References

[1] Sata, M. (2013). Factors affecting consumer buying behavior of mobile phone devices. Mediterranean Journal of Social Sciences, 4(12): 103.

[2] Mihir, H., Patel, M. I., Jani, S., et al. (2021) Diamond price prediction using machine learning, 2021 2nd International Conference on Communication, Computing and Industry 4.0 (C2I4). IEEE, 1-5.

[3] Muhammad, A., & Zafar, Khan. (2018). Mobile Price Class prediction using Machine Learning Techniques, International Journal of Computer Applications, 179(29): 6-11. doi: 10.5120 /ijca2018916555.

[4] Nti, I. K., Adekoya, A. F., Weyori, B. A. (2020) A comprehensive evaluation of ensemble learning for stock-market prediction. Journal of Big Data, 2020, 7(1): 1-40.

[5] Nasser, I. M., Al-Shawwa, M. O., Abu-Naser, S. S (2019) Developing Artificial Neural Network for Predicting Mobile Phone Price Range.

[6] Çetın, M., & Koç. Y. (2021) Mobile Phone Price Class Prediction Using Different Classification Algorithms with Feature Selection and Parameter Optimization, 2021 5th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), Ankara, Turkey, pp. 483-487.

[7] Kaggle (2023) Mobile Phone Price https://www.kaggle.com/datasets/rkiattisak/mobile-phone-price/code

[8] Navid, K., Annan Z., et al. (2021) Improved prediction of slope stability using a hybrid stacking ensemble method based on finite element analysis and field data, Journal of Rock Mechanics and Geotechnical Engineering, Volume 13, Issue 1, Pages 188-201. ISSN 1674-7755.

[9] Luo, M., et al. (2021) Combination of Feature Selection and CatBoost for Prediction: The First Application to the Estimation of Aboveground Biomass, Forests, 12: 216. doi:10.3390/f12020216.

[10] Gao, H., et al. (2019). Application of improved distributed naive Bayesian algorithms in text classification, Journal of Supercomputing, 75: 5831-5847. doi:10.1007/s11227-019-02862-1.

[11] Xu, J., et al. (2018) A survey of ensemble learning approaches (in Chinese). Journal of Yunnan University (Natural Sciences Edition), 40(06): 1082-1092.