Dean&Francis

ISSN 2959-6157

Prediction on Mobile Phone Price Range using Classification Models

Jingyan Lu[1, *]

[1]Department of Statistics, University of Washington, Seattle, 98105, United States

*Corresponding author: Jingyanl@uw.edu

Abstract:

Function and pricing of mobile phones have always been a noticeable question to customers. Countless studies have been done for phone brands to set their standard on pricing new released phones. This paper analyzes the mobile price classification dataset on Kaggle dataset, by using three models: logistic regression, k-nearest neighbors, and support vector machine. Feature selection by considering the correlations between features are used as comparison to the models without feature selection. The performance of models is shown by accuracy and macro F1 score. The performance of models on training dataset and testing dataset are then compared. The conclusion by this paper is that two models of logistic regression have the best performance, followed by two models of k-nearest neighbors which has almost the same performance. Support vector machine with feature selection has a good performance, while it performs poorly without feature selection. It is concluded that feature selection improves the performance of the model significantly. The performance of models on training dataset and testing dataset are consistent.

Keywords: Machine learning; phone price prediction; logistic regression; KNN; SVM.

1. Introduction

Ever since the invention of cellular phones in 1973, they have become more convenient and functional during the past 50 years [1]. Nowadays, mobile phones are considered as necessity due to its convenience in mobile communication and entertainment. Phone brands and different phone models emerged in the market as a result of the growth in need of mobile phones. Due to the difference of requirement to mobile phones and the income level of customers, correctly pricing the phone has become crucial for phone brands [2]. The “brick phones” 50 years ago had almost no functions other than making phone calls, yet they cost about 4000 dollars at that time, which is equivalent to 11500 dollars in 2023. In comparison, an iPhone 16 pro max, a new released smart phone in 2024 with all functions and top-level configuration, are priced 1199 dollars. While time is a considerable factor affecting mobile phone prices, there are other factors influencing price of mobile phones on a horizontal scale. This paper aims to predict phone prices based on various characteristics of phones using machine learning models, making it clearer for people to understand what may lie behind the price of a mobile phone and easier for phone brands to decide the pricing of new released phones.

The most famous dataset used in machine learning of phone price is ‘Mobile Price Classification’ from Kaggle, updated 7 years ago. Many studies on machine learning about phone prices are done using this dataset. Saeed et al. mentioned two ways to predict the price of mobile phones by two different models: Support vector machine (SVM) and Rigid classifier [3]. Support vector machine gives higher accuracy. However, Saeed et al. mentions the overfitting of SVM on the dataset, which makes Rigid classifier a model worth more consideration. Sunariya et al. used 4 other models besides SVM: decision tree, logistic regression, K-nearest neighbor, random forest. In addition, data cleaning was done before prediction. They have reached a result that SVM also performs the best [4]. Güvenç and Koçak compared two models: KNN and Deep Neural Network (DNN) on their performance with the dataset [5]. Chen considerd reducing the number of predictors using two methods: Multilayer Perceptron (MLP) and Principal component analysis (PCA). The features by significant level using two methods were ranked. Accuracy comparison and loss comparison were used to compare the performance of two models under different number of features selected [6]. Çetın also made some feature selection. The methods used were ANOVA and Mutual information. Then the accuracy of Random Forest, Logistic regression, Decision Tree, Linear Discriminant Analysis, KNN and support vector classifier (SVC) with the chosen features are tested. In addition, hyperparameter optimization was performed to optimize the performance of each model. SVC had the highest accuracy after hyperparameter optimization. Çetın then reached a different conclusion with Saeed et.al. in their paper that there is no overfitting in SVC classifier. The process of model selection and hyperparameter could be the reason behind this [7].

Maesya et al. predictd phone price using a different dataset. The phone price of this dataset are continuous variables, different from categorical target in the dataset used by other paper mentioned above. The model they chose are random forest and linear regression. They concluded that random forest performs better [8]. Duan et al. focused on pricing of popular mobile phones in China. They collected the data on a shopping website in China and predicted phone price using SVM [9]. The dataset Parth et.al. used were collected by themselves, which have 24 variables in total. This dataset was collected in 2024, which may indicate that the model used in this paper could have a better performance on predicting the price of new released mobile phones nowadays [10].

To conclude, numerous studies have been made on predicting phone price with various characteristics of a phone. This paper will dive into the performance of SVM, KNN and logistic regression on predicting the price of mobile phone.

2. Methods

2.1 Data Source

The datasets used in this paper is from Kaggle, collected by Abhishek Sharma 7 years ago. One of the datasets is the training dataset, containing information of 2000 phones including the responce variable phone price. The other dataset is the testing dataset with information of 1000 phones, but without phone price.

2.2 Variable Selection

There are a total of 20 features and 1 target variable in the training dataset. Some examples of the features are: battery_power (battery power), blue (with bluetooth or not), px_height (pixel resolution), ram (Random Access Memory in mega bytes). The target variable is price_range with four levels: 0 (low cost), 1 (medium cost), 2 (high cost) and 3 (very high cost). Table 1 and Table 2 show a brief summary of all features, divided into numerical features and categorical features.

Table 1. Summray of numerical features

Features	Mean	Median	Maximum	Minimum
battery power	1238.56	1226	1998	501
clock_speed	1.52	1.5	3	0.5
fc	4.31	3	19	0
int_memory	32.05	32	64	2
m_deap	0.50	0.5	1	0.1
mobile_wt	140.25	141	200	80
n_cores	4.52	4	8	1

pc	9.92	10	20	0
px_height	645.11	564	1960	0
px_width	1251.52	1247	1998	500
ram	2124.21	2146.5	3998	256
sc_h	12.31	12	19	5
sc_w	5.77	5	18	0
talk_time	11.01	11	20	2

Table 2. Summary of categorical features

Features	Number of 0	Number of 1
blue	1010	990
dual_sim	891	1019
four_g	957	1043
three_g	477	1523
touch_screen	994	1006
wifi	986	1014

2.3 Model Selection

This paper uses three machine learning models: logistic regression, k-nearest neighbor (KNN) and Support vector machine (SVM). Logistic regression is a method widely used for predicting discreate outcomes. Instead of predicting the value of response Y, it predicts the probability of response Y belonging to category. For this dataset, the Y to be predicted is price range, which takes 4 values from 0 to 3. It is predicted by the formula:

$$pr\left(Y=k \mid X=x\right)=\frac{e^{\beta_{k0}+\beta_{k1}X_1+\beta_{k2}X_2+\ldots+\beta_{kp}X_p}}{1+\sum_{l=1}^{K-1}e^{\beta_{l0}+\beta_{l1}X_1+\beta_{l2}X_2+\ldots+\beta_{lp}X_p}} \quad (1)$$

Where k is the level, X_i are the predictors and β_i are the corresponding coefficients.

K-nearest neighbor (KNN) is a simple non-parametric method for classification. This model first finds k nearest training data to the test data x, where k is a positive integer to be determined. The distance is usually determined by the Euclidean distance. KNN model then classifies test data x to be the class that most data among the k nearest training data belongs to.

Support vector machine (SVM) is a method suitable for both linear and non-linear models. It aims to find the hyperplane (decision boundary) separates data into different classes the best, so it is often considered as a very effective method.

3. Results and Discussion

3.1 Feature Selection

This paper uses correlation between features and target variable as a way of selecting features for machine learning models. The high correlation between a feature and the target variables means that this feature has a strong relationship with the target, so it would be considered as a useful predictor. The proper selection of predictors has a positive influence on the accuracy of model. Figure 1 shows the correlation between features and target variable.

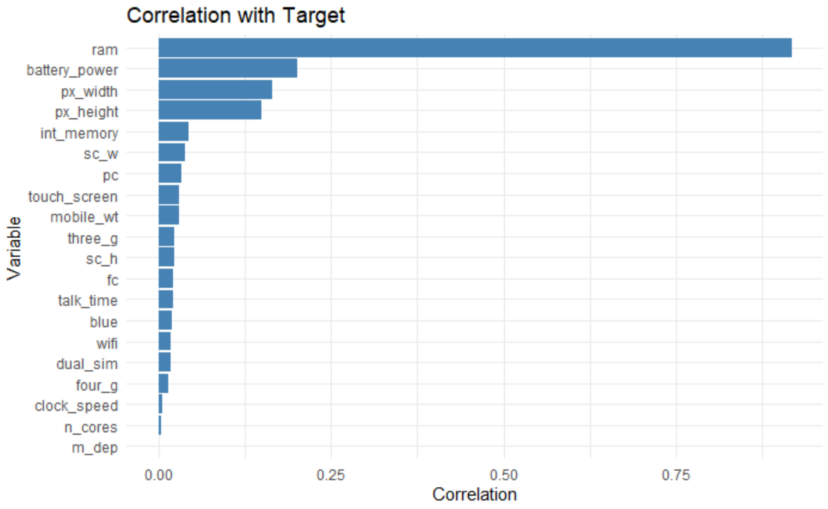

Fig. 1 Correlation between features and target variable

It is shown in figure 1 that ram has a very high correlation with price range, exceeding 0.875, which shows the great importance of ram in the pricing of phone. To elaborate more, figure 2 gives a clear view on how ram is affecting phone price, as the cruves are clearly divided into four parts.

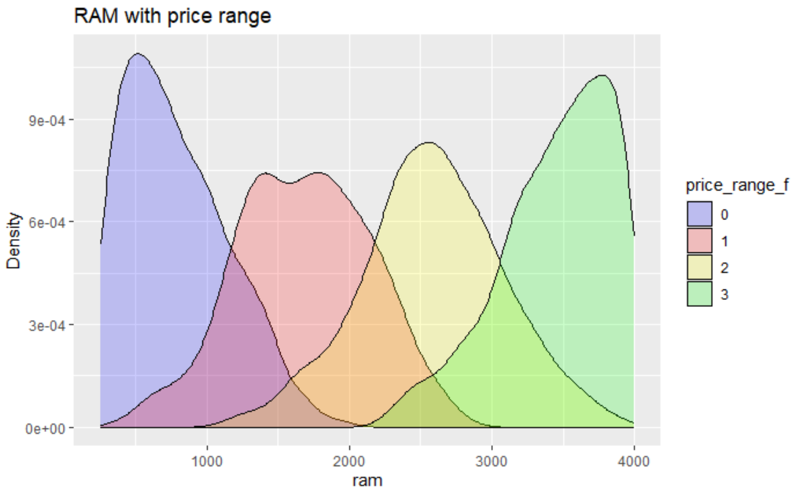

Fig. 2 RAM with price range

Battery_power, px_width and px_height are showing correlation above 0.125, which makes them less crucial compared to ram, but still they have correlation much greater than the rest of the variables. Figure 3 shows the relationship between battery power and price range. It is shown in figure 3 that there is great seperation between price range 0 and 3, but there is a large area of overlapping between 1 and 2.

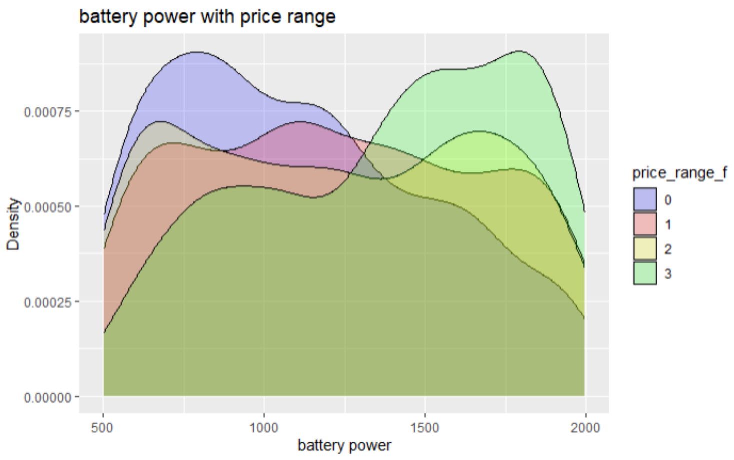

Fig. 3 Battery power with price range

The four variables with highest correlation with price range(ram, battery_power, px_width, px_height) are used in this paper to conduct machine learning models on predicting phone price. Models using all variables as features are also created to compare with the performace of the models using chosen variables.

3.2 Performance of Training Dataset

The models are trained using the training dataset by randomly dividing the data into a ratio of 7:3. Table 3 shows the overall performance of models using accuracy and macro F1 score. Accuracy is calculated by:

$$Accuracy = \frac{NumberofCorrectpredictions}{Totalnumberofpredictions} = \frac{TP+TN}{TP+TN+FP+FN} \qquad (2)$$

Where TP: true positives, TN: true negatives, FP: false positives, FN: false negatives. Macro F1 score is often used to show the performance of multiclass classification, which is calculated by the average score of F1 score of all classes.

Table 3. Comparison of performance on each model

Models	Accuracy	Macro F1 score
Logistic regression(all)	0.9775	0.9772
Logistic regression(selected)	0.9595	0.9592
SVM (all)	0.8589	0.8580
SVM (selected)	0.9520	0.9516
KNN (all)	0.9354	0.9346
KNN (selected)	0.9459	0.9451

* All: model uses all predictors, selected: model only considers the 4 predictors mentioned above

In KNN models, k that would give the highest accuracy are chosen. The k used for KNN model without feature selection is 13, and with feature selection is 14. All models are showing great performance of accuracy and Macro F1 score over 90% except for SVM without feature selection. The model with best performance is logistic regression without feature selection. This model is with accuracy and macro F1 score both over 97%. The model that shows worse performance is SVM model without feature selection, giving an accuracy and macro F1 score of 86%.

In logistic regression and KNN models, the difference between models with feature selection and without feature selection is not significant. There is a -1.8% difference in accuracy and Macro F1 score in two logistic regression models. As in KNN models, there is a 1.1% difference in accuracy and macro F1 score between two models of KNN. However, in SVM, the difference in accuracy and macro F1 score is very significant compared to other two type of models. The accuracy and macro F1 score for SVM model with feature selection improved by 9.3% and 9.4% compared to SVM model without feature selection.

3.3 Performance of Testing Dataset

The testing dataset is without price range, so it is not possible to test the accuracy and Macro F1 score of the models (Table 4). What is done instead is comparing the difference in predictions between the same model with different selection of variables. The difference rate is calculated by:

$$Differencerate = \frac{Numberofdifferentpredictions}{Totalnumberofpredictions} \quad (3)$$

Table 4. Comparison of perfermance between same type of model on training dataset

Models	Difference rate
Logistic Regression	0.046
SVM	0.133
KNN	0.027

It could be seen from the table that two models of KNN has the lowest difference rate 0.027, followed by logistic regression with difference rate 0.046. SVM has the highest difference rate of 0.133. The performance of models on testing dataset is consistent with the performance of models on training dataset. The accuracy and Macro F1 score of logistic regression and KNN does not differ significantly with and without model selection on the training dataset, which remains the same when testing on the testing dataset. For SVM, the difference between models are both high on training and testing dataset.

4. Conclusion

Considering the performances of different models mentioned in this paper, it could be concluded that logistic regression without feature selection has the best performance on predicting the price range of mobile phones. The impact of feature selection on the performance of models are not significant in logistic regression and KNN. For SVM, the model with feature selection shows a better performance. The discrepancy in performance of feature selection on different models could be caused by the characteristic and difference in computation method of different models. Another reason for the discrepancy might be the randomness in split of dataset, which may cause some models to perform uncommonly well or bad.

Some suggestions can be provided to future studies of predicting phone price level based on this paper. As shown in variable selection, about half of the phones are without touch screen, Bluetooth, wifi and four-G, which does not accord with the features of mobile phones nowadays. Other later datasets of phone price would be a better choice for studying on building models for phone price if the models are used for phone price predictions on new released phones. Alternative ways of feature selection are also worth considering in future studies to improve the accuracy of models.

References

[1] Jessop G. A brief history of mobile telephony: The story of phones and cars. Southern Review: Communication, Politics & Culture, 2006, 38(3): 43-60.

[2] Alfred O. Influences of price and quality on consumer purchase of mobile phone in the Kumasi Metropolis in Ghana a comparative study. European Journal of Business and Management, 2013, 5(1): 179-198.

[3] Saeed A, Mukhtar A, Arafat Y, Abbas M, Saeed A. Intelligent Assessment of Secondhand Mobile Phone Prices by Machine Learning Techniques. Journal of Computing & Biomedical Informatics, 2024.

[4] Sunariya N, Singh A, Alam M, Gaur V. Classification of Mobile Price Using Machine Learning. Working paper, 2024, 55-66.

[5] Güvenç E, Çetin G, Koçak H. Comparison of KNN and DNN classifiers performance in predicting mobile phone price ranges. Advances in Artificial Intelligence Research, 2021, 1(1): 19-28.

[6] Chen M. Mobile Phone Price Prediction with Feature Reduction. Highlights in Science, Engineering and Technology, 2023, 34: 155-162.

[7] Çetın M, Koç Y. Mobile phone price class prediction using different classification algorithms with feature selection and parameter optimization. In 2021 5th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), 2021, 483-487.

[8] Maesya A, Yanfi Y. Mobile Phone Price Prediction Based on Supervised Learning Algorithms. International Journal of Applied Engineering and Technology, 2023, 5(1): 41-44.

[9] Duan Z, Liu Y, Huang K. Mobile phone sales forecast based on support vector machine. In Journal of Physics: Conference Series, 2019, 1: 12061.

[10] Bhatnagar P, Lokesh G H, Shreyas J, Flammini F, Panwar D, Shree S. Prediction of Mobile Phone Prices using Machine Learning. In Proceedings of the 2024 9th International Conference on Machine Learning Technologies, 2024, 6-10.