Using Machine Learning Methods to Predict Mobile Phone Prices

Wenbo Lu

Mathematical and Statistical Data Science, Faculty of Science, University of Technology Sydney, Sydney, New South Wales, Australia
Wales-Australia-Wenbo.Lu@student.uts.edu.au

Abstract. This study analyzes the price range and configurations of mobile phones on the market to help emerging mobile phone companies accurately position their product prices and better compete in the mobile phone market. The study uses machine learning techniques to predict mobile phone prices, employing three methods: logistic regression, random forest, and Multilayer Perceptron (MLP) to build models, and generating prediction accuracy and feature importance maps related to mobile phones for each model. Through comparison of model accuracy, this study found that the logistic regression model performed best, with the highest prediction rate. Analysis of the feature importance maps and exploratory analysis revealed that feature combinations have a more significant impact on pricing than single features. In addition, an important finding is that higher-priced mobile phones are less likely to support memory expansion. This study not only achieves price prediction but also provides an important reference for the research and development and marketing of electronic products.

Keywords: Logistic Regression, Random Forest, MLP.

1 Introduction

With the development of society, mobile phones have become an indispensable part of people's daily lives, and the mobile phone industry has become one of the most prosperous electronic consumer industries in the world. For companies that want to enter the mobile phone market, how to price-position their mobile phone products has become a difficult problem. By analyzing the best-selling mobile phones on the market, finding the important factors related to price positioning has become the primary goal of these emerging companies. This study divides mobile phones of different prices on the market into four price positioning categories to facilitate modeling. This choice has also been recognized by many studies in the same field [1,2]. This study mainly constructs three models: logistic regression, random forest and Multilayer Perceptron (MLP) to achieve the purpose. Among them, the logistic regression model has strong interpretability and can intuitively show the relationship between mobile phone configuration and price positioning [3,4]. The random forest model can more clearly grasp the nonlinear relationship between features and find important features related to

K. Subramanian (ed.), *Proceedings of the International Workshop on Advances in Deep Learning for Image Analysis and Computer Vision (IWADIC 2025)*, Advances in Computer Science Research 128,
https://doi.org/10.2991/978-94-6239-648-7_71

price positioning [5,6]. The MLP model captures more complex feature interactions through linear changes and nonlinear activations, thereby showing the superposition of multiple feature interactions [7]. The three models can be used together to produce complementary results and better capture those features related to price [8]. In the process, this study uses VIF to suppress the multicollinearity problem [9]. This study uses accuracy as the main result, and adds precision, recall, F1, etc., to determine which model has stronger predictive ability. And by replacing importance and coefficient strength, a feature importance graph is generated to show the features that are strongly correlated with price [10]. Some similar studies only use accuracy as the result, without considering more indicators, and thus cannot fully reflect the performance of the model. In order to solve this problem, this study adds precision, recall, F1, etc., on the basis of accuracy to fully demonstrate the model. And through these results and feature importance graphs, explore the features that are strongly correlated with the price of mobile phones.

This study is mainly divided into five parts. The first part is "Introduction", which explains the research background, purpose and concepts. The second part is "Dataset", which includes the source, description, preprocessing and exploratory analysis of the dataset. The third part is "Model", which describes the principles of the three models used. The fourth part is "Results", which achieves the purpose by analyzing accuracy and feature importance. The last part is "Conclusion", which summarizes the research and report.

2 Dataset

2.1 Descriptive Introduction to the Dataset

The data for this study comes from the Kaggle platform and contains 866 rows of data and 26 feature variables, totaling 22,062 data points, with "price" as the target variable. There are 18 categorical features, such as "num_rear_cameras", "has_ir_blaster", "fast_charging_available", and "num_front_cameras", totaling 15,236 data points. There are 8 numerical features, such as "processor_speed", "battery_capacity", "fast_charging", and "screen_size", totaling 6,826 data points. There are 29 brands represented, with Samsung, Realme, Xiaomi, and Apple having the most.

2.2 Data Preprocessing

Data preprocessing was performed on the dataset. Five features irrelevant to phone performance were removed during preprocessing: "brand_name", "model", "processor_name", "processor_brand", and "rating". Additionally, "extended_upto" was also removed due to its large number of missing values.The dataset was divided into a 75% training set and a 25% test set, with a "split" function added for differentiation.

Since some features contained numerous and complex values, binning was performed to simplify the model.

The "price" data (currency unit: Indian Rupee) is divided into "0", "1", "2", and "3". The range for "0" is 2799 to 10423, for "1" it is 10424 to 16999, for "2" it is 17000 to

27999, and for "3" it is 28000 and above. The quantities at each of the four price levels are similar, indicating an overall balance.

The resolution in "resolution" is divided into six levels from 1" to "6" according to the total pixel tree (MPx). According to the conversion formula

$$MPx = \frac{widt \times eig \quad t}{1{,}000{,}000} \tag{1}$$

The range of "1" is MPx < 0.7, the range of "2" is 0.7 ≤ MPx < 1.0, the range of "3" is 1.0 ≤MPx < 1.6, the range of "4" is 1.6 ≤ MPx < 2.3, the range of "5" is 2.3 ≤ MPx < 3.7, and the range of "5" is MPx ≥ 3.7. Generate a new feature "resolution_level" and add it to the dataset.

Based on battery capacity (unit: mAh), "battery_capacity" is divided into 6 levels (1-6): 1 represents less than 2500, 2 represents 2500 to 3499, 3 represents 3500 to 4499, 4 represents 4500 to 5499, 5 represents 5500 to 6499, and 6 represents greater than 6500. A new feature "battery capacity level" is generated and added to the dataset.

Based on clock speed (GHz), the "processor_speed" field is divided into six levels, from "1" to "6". "1" is less than 1.4, "2" is 1.4 to 1.79, "3" is 1.8 to 2.19, "4" is 2.2 to 2.59, "5" is 2.6 to 2.99, and "6" is greater than or equal to 3.0. "processor_speed_level" has been added to the dataset.

The "fast charging" option is divided into six levels based on charging efficiency (in watts), from "1" to "6". "1" is 10 to 17 watts, "2" is 18 to 29 watts, "3" is 30 to 44 watts, "4" is 45 to 64 watts, "5" is 65 to 99 watts, and "6" is ≥100 watts. This setting directly replaces the original "fast_charging" option.

The "price" is divided into four price levels (as shown in Figure 1). The number of phones in each price level is similar, and the overall performance is quite flat. In the exploratory analysis and model building process, "battery_capacity_level" "processor_speed_level" and "resolution_level" replaced "battery_capacity" "processor_speed" and "resolution".

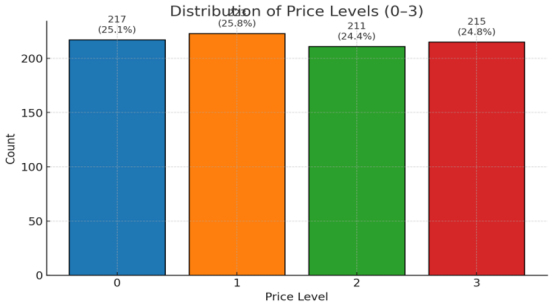

Fig. 1. Division of target (Picture credit: Original)

2.3 Exploratory Analysis (EDA)

Exploratory analysis was employed to analyze the data, focusing on binary features, numerical features, and all features, to analyze the key characteristics of mobile phone

prices and their correlations. The core variable in this analysis was "price," including price levels 0, 1, 2, and 3.

The binary features include “has_5g”, “has_nfc”, “has_ir_blaster”, “fast_charging_available”, “os”, and “extended_memory_available”, which represent the distribution of binary features at different price points in the dataset.

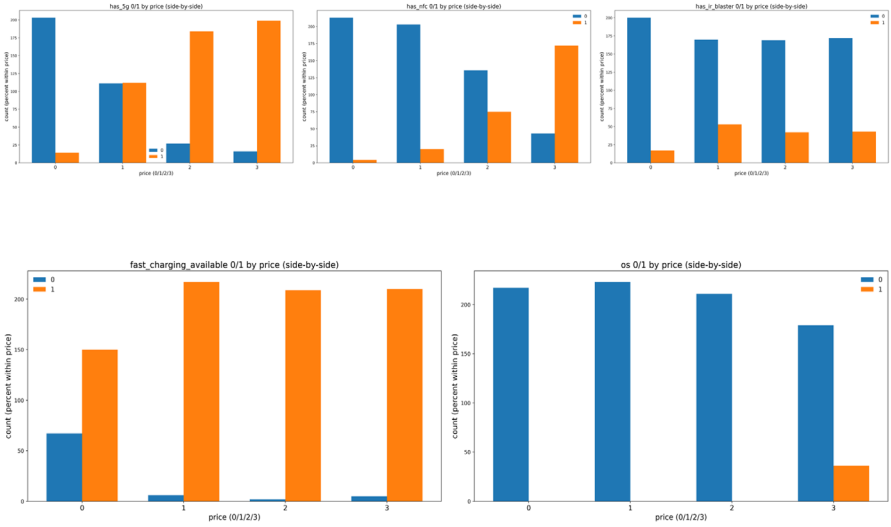

Fig. 2. Number of binary features at different price levels (Picture credit: Original)

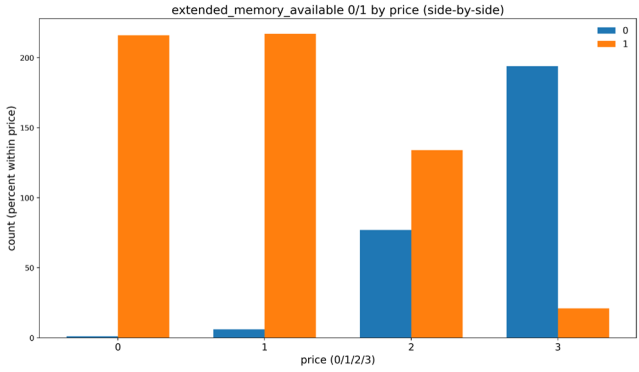

Fig. 3. Number of binary features at different price points (Picture credit: Original)

As shown in Figures 2 and 3, higher-priced phones are more likely to support 5G, NFC, fast charging, and infrared blasters. Regarding operating systems, Android is widely available across all price points, while iOS tends to be found in higher-priced phones. Finally, higher-priced phones are less likely to support memory expansion.

The numerical features include "battery_capacity_level", "processor_speed_level", "resolution_level", "ram_capacity", "internal_memory", "screen_size", "refresh_rate", "nums_core", "num_rear_cameras", "num_front_cameras", "primary_camera_rear", "primary_camera_front", and "fast_charging".

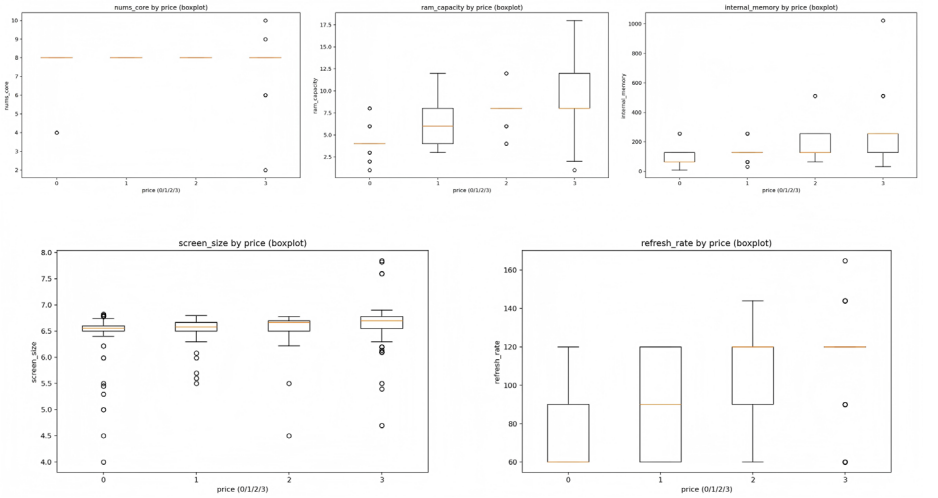

Fig. 4. Numerical feature distribution at different price points1 (Picture credit: Original)

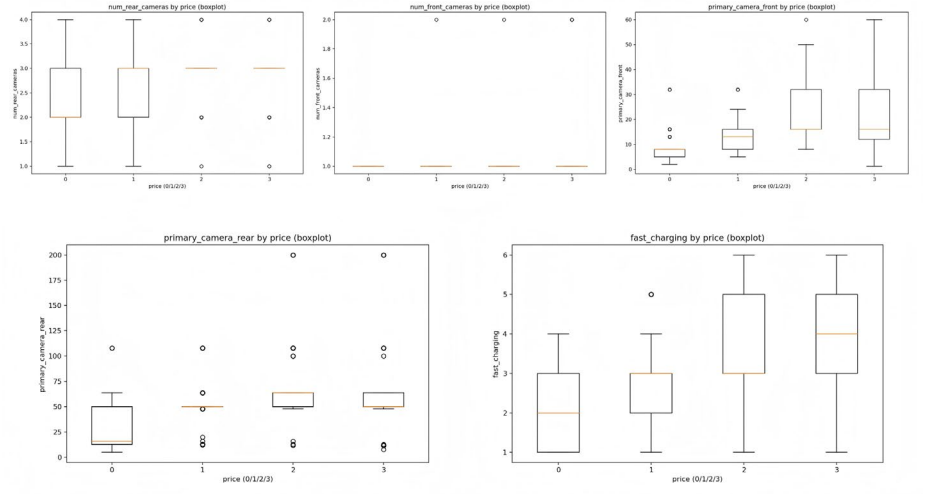

Fig. 5. Numerical feature distribution at different price points2 (Picture credit: Original)

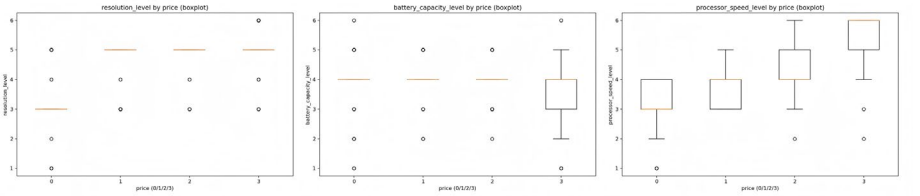

Fig. 6. Numerical feature distribution at different price points3 (Picture credit: Original)

A comprehensive analysis of Figures 4, 5, 6 clearly shows that, with the exception of the number of processor cores, number of front-facing cameras, screen size, and battery capacity, which remain similar across all price points, all other features increase with price. RAM, internal memory, screen refresh rate, rear camera pixel count, fast charging wattage, resolution, and processor clock speed all show a significant positive correlation with price. Finally, this paper analyze the correlation heatmaps for all features. Figure 7 shows strong correlations between phone configurations (that is, individual features). For example, "fast_charging_available" and "fast_charging" are strongly correlated, as are "ram_capacity" and "internal_memory." This indicates multicollinearity in the dataset, which requires careful consideration in the subsequent linear modeling. In addition, by observing and analyzing the correlation heat map, this paper can intuitively see the characteristics of significant positive and negative correlation with price, and ultimately obtain results similar to the previous box plot analysis.

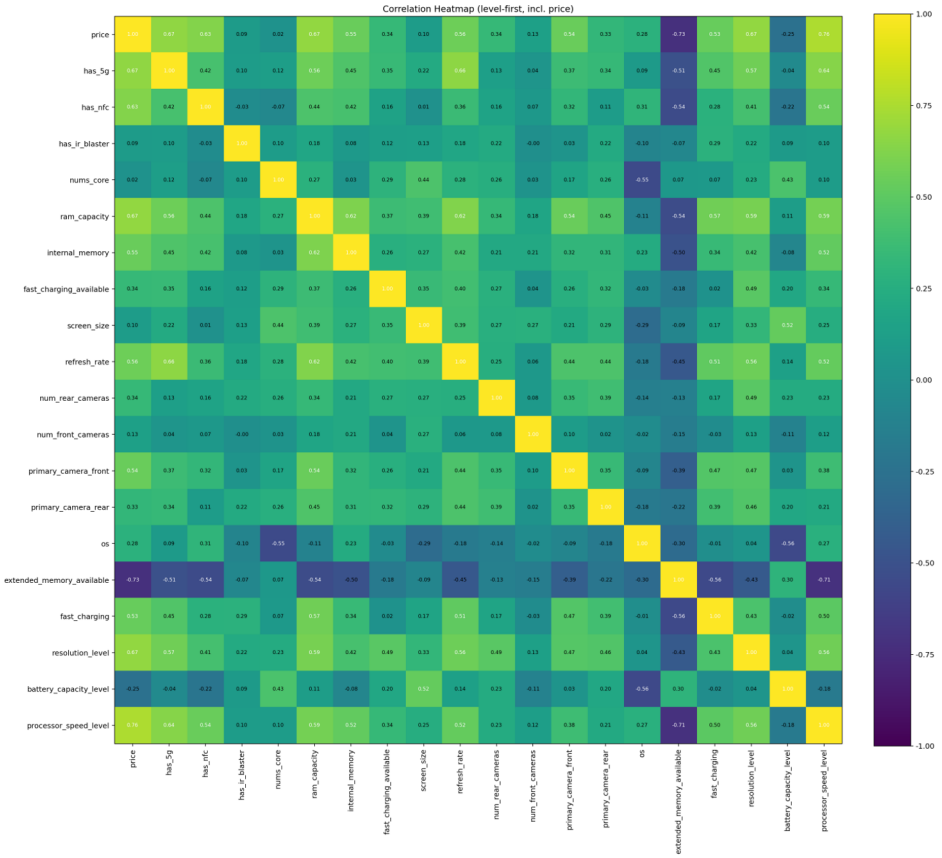

Fig. 7. Correlation heatmap of all features (Picture credit: Original)

3 Model

A total of three models were constructed in this study, namely the logistic regression model, the random forest model and the MLP model.

3.1 Logistic Regression Model

Logistic regression is one of the most common linear classification models, often used for binary classification. The advantages of this type of model are its interpretability, fast training speed, and robustness.

Logistic regression models perform linear scoring on features

$$z = \beta^{\mathrm{T}} x \tag{2}$$

Convert the feature into a score that can be understood and explained, and then pass it through the binary classification output function (sigmoid)

$$P(y = 1|x) = \sigma(z) = \frac{1}{1 + e^{-z}} \tag{3}$$

By mapping the scores to the (0, 1) interval, the function output, P, can be directly interpreted as a probability.

The target feature "price" in this study is a four-category feature. Therefore, in the current logistic regression model, the binary output function "sigmoid" needs to be generalized to the multi-category output function "softmax"

$$softmax(z)_k = \frac{e^{z_k}}{\sum_{j=1}^{K} e^{z_j}}, k = 1, \ldots, K \tag{4}$$

This paper can achieve the effect of logistic regression in a binary classification model. On this basis, this study add L2 regularization to become a multinomial logistic regression with L2 regularization

$$L_{(\beta)} = -\sum_{i=1}^{n}\sum_{k=1}^{K} y_{ik} log \frac{e^{\beta_k^T x_i}}{\sum_{j=1}^{K} e^{\beta_j^T x_i}} + \frac{\lambda}{2}\sum_{k=1}^{K} (\| \beta_k \|_2)^2 \tag{5}$$

Implement shrinkage coefficients to prevent individual coefficients from increasing due to strong correlations in the data set, and also use VIF (variance inflation factor)

$$VIF_j = \frac{1}{1 - R_j^2} \tag{6}$$

Further mitigate the multicollinearity in the dataset. Finally, a robust and effective model is generated. β refers to the coefficient vector (parameter), x refers to the eigenvector, z refers to linear score, $\beta^{\mathrm{T}} x$ refers to the linear score (weighted sum of features), $\sigma(z)$ refers to the sigmoid activation function.

3.2 Random Forest Model

Compared to logistic regression, random forest improves the overall performance of the model by forming multiple decision trees through ensemble learning. Each tree uses different samples and feature subsets, and finally obtains the final prediction result through majority voting or averaging of all trees. This enables random forest to better capture the nonlinear relationship between features, is insensitive to feature scaling, and is more robust to multicollinearity. In this process, a single tree follows the Gini impurity.

$$G(S) = 1 - \sum_{k=1}^{K} p_k^2 \tag{7}$$

During the splitting process.

This makes each tree more concentrated after splitting, achieving smaller classification errors and better probability estimation. Finally, the random forest is used to make the final prediction get the result.

$$\hat{y}(x) = arg\ max_k \sum_{t=1}^{T} 1\{f_t(x) = k\} \tag{8}$$

p_k refers to the proportion of the sample set belonging to category k, T refers to the number of decision trees in the forest.

3.3 MLP Model

MLP has the ability to capture both nonlinear relationships and high-order interactions, and can characterize the combined effects of multiple features.

Unlike the previous two models, the MLP model is primarily structured into three layers: the input layer, the hidden layer, and the output layer. The input layer receives preprocessed features and generates a separate channel for each feature. The hidden layer is automatically determined through hyperparameter search and may be one or two layers. If there is only one layer, it may have 256 or 128 neurons. If there are two layers, the candidates are (256, 128) or (128, 64). The hidden layer performs a linear transformation

$$z = w^{\top}x + b \tag{9}$$

on each layer. After performing weighted summation on the output of the previous layer, nonlinear activation

$$a = \varphi(z) \tag{10}$$

is performed. The MLP model can capture and display nonlinear relationships between features.

Finally, to output the results to the "output layer", a multi-class output function is used.

$$softmax(z)_k = \frac{e^{z_k}}{\sum_{j=1}^{K} e^{z_j}}, k = 1, \ldots, K \tag{11}$$

In this process, L2 regularization is also used to suppress weights, reduce overfitting, and improve generalization. w refers to the weight vector of the current layer, x refers to the output (or original input) of the previous layer, b refers to the bias term, z refers to linear scoring, $\varphi(z)$ refers to the nonlinear activation function.

4 Experiment

4.1 Experimental Configuration

This research was conducted using Google Colab, using Python 3.10 as the programming language.

Logistic regression, random forest, and MLP models share the same evaluation metric: accuracy. Higher accuracy indicates stronger model recognition capabilities. Accuracy is also reported along with recall, precision, F1 score, micro-average, macro-

average, and weighted average. Finally, a feature importance plot is generated, showing features that are positively correlated with price.

4.2 Experimental Results and Analysis

Table 1. Results comparison

Model	Logistic regression	Random Forest	MLP
Accuracy	0.7926	0.788	0.714
Macro Precision	0.7915	0.784	0.715
Macro Recall	0.7927	0.789	0.713
Macro F1	0.7919	0.783	0.709
Weighted Precision	0.7917	0.785	0.715
Weighted Recall	0.7926	0.788	0.714
Weighted F1	0.7919	0.783	0.710
Micro Precision	0.7926	0.788	0.714
Micro Recall	0.7926	0.788	0.714
Micro F1	0.7926	0.788	0.714

Table 1 shows the experiment results. This study constructed three models: logistic regression, random forest, and MLP. The logistic regression model achieved an accuracy of 0.793, surpassing both the random forest and MLP models. Furthermore, logistic regression outperformed both the random forest and MLP models in precision, recall, and F1. Higher accuracy indicates better overall predictive ability, while higher precision and recall indicate fewer false negatives and underestimations. A higher F1 indicates a strong balance between precision and recall. The results demonstrate that the logistic regression model offers superior overall performance.

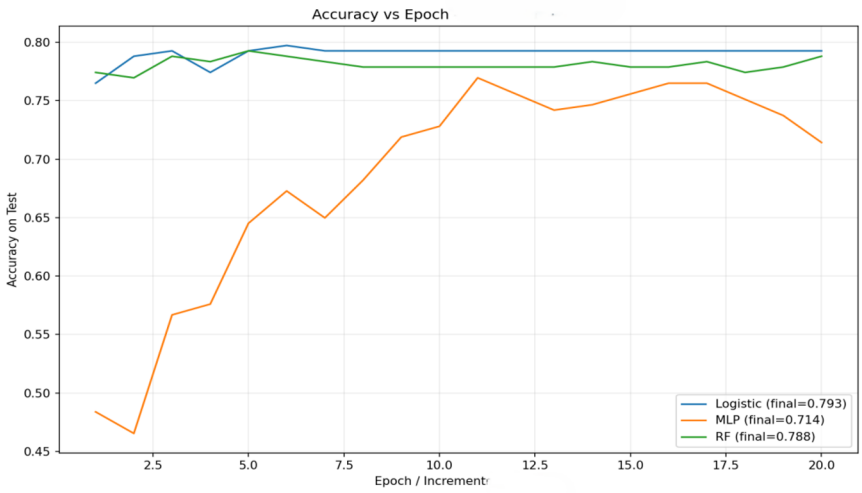

Fig. 8. Accuracy change curve (Picture credit: Original)

Figure 8 shows the accuracy curves of the three models. The MLP model represents a true epoch, the random forest model simulates an epoch by increasing the number of trees, and the logistic regression model simulates an epoch by iterating the interpreter and sweeping the regularization strength.

As shown in the figure, the accuracy of the logistic regression model remains between 0.75 and 0.8, and stabilizes with increasing epochs. This demonstrates that the logistic regression method effectively processes features and is robust.

The accuracy of the logistic regression model consistently remained between 0.75 and 0.8, and tended to stabilize with increasing epochs, demonstrating the best performance.

The accuracy of the random forest model was generally lower than that of the logistic regression model, and its accuracy decreased with increasing training epochs, indicating that the non-linear relationships between features were not significant.

The accuracy curve of the MLP model initially increases with increasing epochs, then eventually falls back to 0.714. This pattern of the MLP accuracy curve follows the pattern of learning patterns between data and then understanding the uniqueness of each feature. The initial high and then low accuracy curve also indicates the need for stronger regularization to achieve optimal output results. From the comprehensive accuracy change chart, it can be seen that the simpler and more intuitive logistic regression model is more suitable for this data.

4.3 Feature Importance Analysis

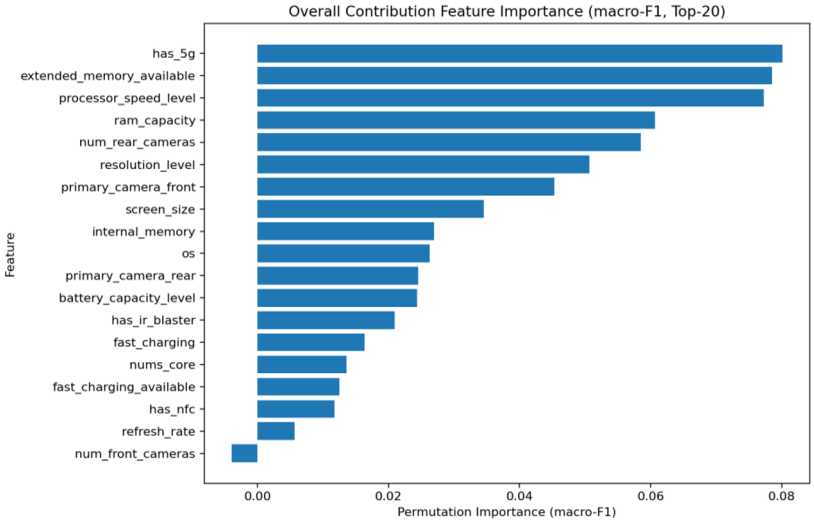

Fig. 9. Logistic regression feature importance graph (Picture credit: Original)

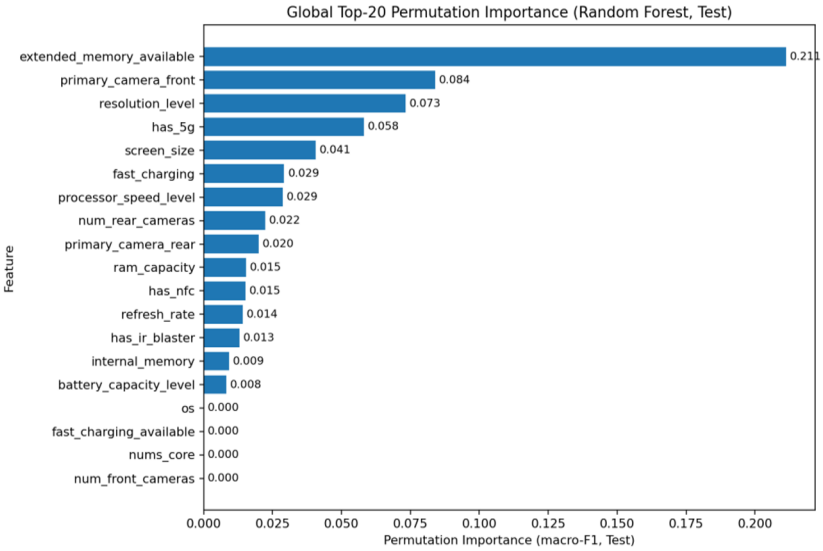

Fig. 10. Random Forest Feature Importance Graph (Picture credit: Original)

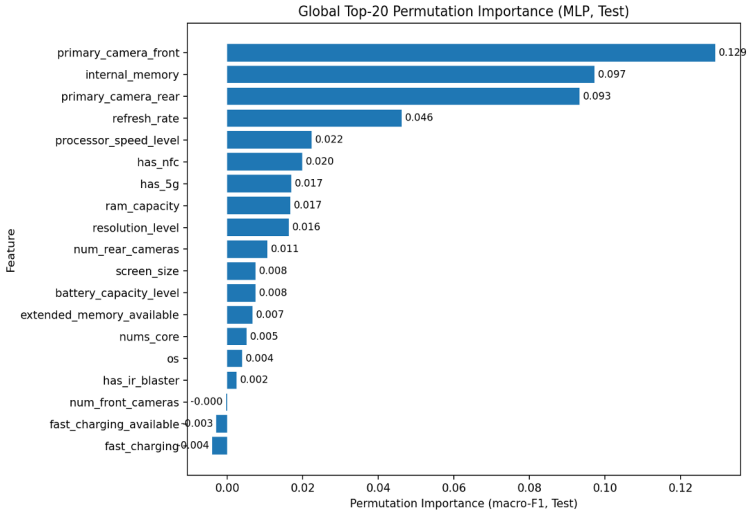

Fig. 11. MLP feature importance graph (Picture credit: Original)

Figures 9, 10, and 11 show the feature importance of the logistic regression, random forest, and MLP models.

In the logistic regression model, "has_5g," "extended_memory_available," and "processor_speed_level" are more important, indicating that the logistic regression model prioritizes strong signals that can be linearly segmented.

In the random forest model, "extended_memory_available" is far more important than other features, indicating that support for memory expansion is a key factor in differentiating price points and demonstrating the random forest's sensitivity to features with clear thresholds.

In the MLP model, the most important feature is "primary_camera_front," while "internal," "primary_camera_rear," and "refresh_rate" also have significant effects. This indicates that the MLP model prioritizes features that can be combined to produce an effect.

The results of the three models show that features such as image display and memory processing have a greater impact. The analysis suggests that when predicting mobile phone prices, attention should be paid to features that can be combined and their combined effects, rather than focusing excessively on individual features.

5 Conclusion

This study constructed logistic regression, random forest, and MLP models. The results of these three models help emerging mobile phone companies better position the price of their new products.

Feature importance analysis of the three models and EDA results show that the higher the overall configuration, the higher the price of the mobile phone, and the worse the memory expansion capability.

Logistic regression demonstrated the strongest predictive power. Random forest ranked second, with its "available expandable memory" feature significantly improving predictive performance. MLP models performed moderately. This indicates a stronger linear relationship between the data.

Logistic regression visually presents key features, random forest models identify the most influential features, and MLP demonstrates the degree of influence of features. These three models can be combined to obtain more comprehensive analytical results.

The limitations of this study are the relatively small sample size, the lack of visualization of the strong correlations between features, and the significant linear relationship in the processed data, all of which affected the research results.

The research findings can also provide insights for electronic product marketing strategies. When developing and promoting new products, it is essential to highlight core selling points while also focusing on improving the overall configuration of the phone.

References

1. Çetın, M., Koç, Y.: Mobile Phone Price Class Prediction Using Different Classification Algorithms with Feature Selection and Parameter Optimization. In: Proc. 5th Int. Symp. Multidisciplinary Studies and Innovative Technologies (ISMSIT), Ankara, Turkey, pp. 483–487 (2021)

2. Sunariya, N., Singh, A., Alam, M., Gaur, V.: Classification of Mobile Price Using Machine Learning. In: Proc. Symposium on Computing & Intelligent Systems (SCI), New Delhi, India, pp. 55–66 (2024)
3. Kalaivani, K.S., Priyadharshini, N., Nivedhashri, S., Nandhini, R.: Predicting the Price Range of Mobile Phones Using Machine Learning Techniques. In: AIP Conference Proceedings, vol. 2387, no. 1, p. 140010. AIP Publishing LLC (2021)
4. Sperandei, S.: Understanding Logistic Regression Analysis. Biochemia Medica 24(1), 12–18 (2014)
5. Reddy, M.D.K., Geetha, R.: A Novel Calculation Approach for Price Range Prediction with General Optimization of Features Using Deep Neural Network and Random Forest Algorithms. In: AIP Conference Proceedings, vol. 2853, no. 1, p. 020002. AIP Publishing LLC (2024)
6. Liu, Y., Wang, Y., Zhang, J.: New Machine Learning Algorithm: Random Forest. In: Proc. Int. Conf. Information Computing and Applications, Berlin, Heidelberg: Springer, pp. 246–252 (2012)
7. Taud, H., Mas, J.F.: Multilayer Perceptron (MLP). In: Geomatic Approaches for Modeling Land Change Scenarios, Cham, Switzerland: Springer, pp. 451–455 (2017)
8. Arévalo-Cordovilla, F.E., Peña, M.: Comparative Analysis of Machine Learning Models for Predicting Student Success in Online Programming Courses: A Study Based on LMS Data and External Factors. Mathematics 12(20), 3272 (2024)
9. Miles, J.: Tolerance and Variance Inflation Factor. Wiley StatsRef: Statistics Reference Online (2014)
10. Schmidt, F.: The Importance of Replacement Value. Accounting Review, 235–242 (1930)