Smart Mobile Phone Price Prediction using Machine Learning

Author : Preeti
Team Name : Peeku vision

Date of submission
15 July 2023

ABSTRACT

To predict “If the mobile with given features will be Economical or Expensive” is the main motive of this research work. Real Dataset is collected from website www.kaggle.com. Different feature selection algorithms are used to identify and remove less important and redundant features and have minimum computational complexity. Different classifiers are used to achieve as higher accuracy as possible. Results are compared in terms of highest accuracy achieved and minimum features selected. Conclusion is made on the base of best feature selection algorithm and for the given dataset. This work can be used in any type of marketing and business to find optimal product (with minimum cost and maximum features). Future work is suggested to extend this research and find more sophisticated solution to the given problem and more accurate tool for price estimation.

1. INTRODUCTION

Price is the most effective attribute of marketing and business. The very first question of costumer is about the price of items. All the costumers are first worried and thinks “If he would be able to purchase something with given specifications or not”. So to estimate price at home is the basic purpose of the work. This paper is only the first step toward the above mentioned destination.

Artificial Intelligence-which makes machine capable to answer the questions intelligently- now a days is very vast engineering field. Machine learning provides us best techniques for artificial intelligence like classification, regression, supervised learning and unsupervised learning and many more. Different tools are available for machine learning tasks like MATLAB, Python, etc. We can use any of classifiers like Decision tree, Naïve Bayes and many more. Different type of feature selection algorithms are available to select only best features and minimize dataset. This will reduce computational complexity of the problem. As this is optimization problem so many optimization techniques are also used to reduce dimensionality of the dataset.

Smart Mobile Phones, now a days is one of the most selling and purchasing device. Every day new smart mobile phones with new version and more features are launched. Hundreds and thousands of mobiles are sold and purchased on daily basis. So here the mobile price prediction is a case study for the given type of problem i.e.finding optimal product. The same work can be done to estimate real price of all products like cars, bikes, generators, motors, food items, medicine etc.

Many features are very important to be considered to estimate price of mobile. For example Processor, camera, weight of the mobile. Battery Power is also very important in today’s busy schedule of human being. Size and thickness of the mobile are also important decision factors. Internal memory, Camera pixels, and video quality must be under consideration. Internet browsing is also one of the most important constraints in this technological era of 21st century. And so the list of many features based upon those, mobile price is decided. So we will use many of above mentioned features to classify whether the mobile would be very economical, economical, expensive or very expensive.

The structure of the paper is as follows. Next section is review of related work.3rd Section contains Methodology and Experimental procedure. Section 4 is the summary of the results. Comparative study is done in section 5. After that paper is concluded in section 6. Outcomes of the work are discussed in section 7. At last in 8th section some suggestions about future work are given.

2. RELATED WORK

Using previous data to predict price of available and new launching product is an interesting research background for machine learning researchers. Sameer Chand-Pudaruth[1] predict the prices of second hand cars in Mauritius. He implemented many techniques like Multiple linear regression, k-nearest neighbors (KNN), Decision Tree, and Naïve Bayes to predict the prices. Sameer Chand-Pudaruth got Comparable results from all these techniques. During research it was found that most popular algorithms i.e. Decision Tree and Naïve Bayes are unable to handle, classify and predict Numerical values. Number of instances for his research was only 97(47 Toyota+38 Nissan+12 Honda). Due to less number of instances used, very poor prediction accuracies were recorded.

Shonda Kuiper [2] has also worked in the same field. Kuiper used multivariate regression model to predict price of 2005 General Motor cars. He collected the data from available online source www.pakwheels.com. The main part of this research work is “Introduction of suitable variable selection techniques, which helped to find that which variables are more suitable and relevant for inclusion in model. This (His research) helps students and future researchers in many fields to understand the conditions under which studies should be conducted and gives them the knowledge to discern when appropriate techniques should be used.

Support Vector Machine (SVM) concept is used by one another researcher Mariana Listiani [3] for the same work. Listiani predicted prices of leased cars using above mentioned technique. It was found in this research that SVM technique is far more better and accurate for price prediction as compared to other like multiple linear regression when a very large data set is available. The researcher also showed that SVM also handles high dimensional data better and avoids both the under-fitting and over-fitting issues. To find important features for SVM Listiani used Genetic Algorithm. However, the technique failed to show in terms of variance and mean standard deviation why SVM is better than simple multiple regression.

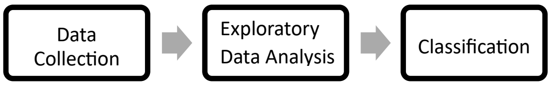
Neural Networks (NN) are more better in estimating price of house, this was concluded in the research of Limsombunchai

[4]. By comparing with hedonic method his method was more accurate Operation of both the methods are same, but in NN the model is trained first and then tested for prediction. Using both the methods NN produced higher R-sg and smaller root mean square error (RMSE), while hedonic produced lower values. This research was limited because the actual house price were missing and only estimated prices were used for the research work.

K Noor and Saddaqat J [5] also worked to predict the price of Vehicles using different techniques. The researchers achieved highest accuracy using multiple linear regression. This paper proposes a system where price is dependent variable which is predicted, and this price is derived from factors like vehicle's model, make, city, version, color, mileage, alloy rims and power steering.

3. METHODOLOGY

The experiment is performed using Google Collab. The main steps of machine learning model are as follows



3.1 Data Collection

The features of mobile phone are collected from www.kaggle.com i.e.

Category (whether the given mobile is made by Apple, Samsung, Lenovo, NOKIA etc). **Dual sim slot, 4G/5G support** and **Bluetooth** is considered as feature whether it is present or not.

Size of display (cm), **weight**(g), **Thickness** (mm), **Internal memory size** (GB), **Camera Pixels** (MP), **Micro processor clock speed**, **RAM size**(GB) and **Battery** (mAh) , **Number of cores in processor etc.** all these attributes have real values with following distinctions.

Table 1. Dataset distinct values

Features	Minimum	Maximum	Mean	StdDiv
Battery power (mAh)	500	1999	1.25	432
Microprocessor Clock speed	0.5	3	1.54	0.83
Camera(MP)	0	19	4.59	4.46
Internal memory(GB)	2	64	33.7	18.1
Mobile depth (cm)	0.1	1	0.52	0.28
Mobile weight	80	200	140	34.8
Number of cores	1	8	4.52	2.29
Bluetooth	0	1	0.49	0.5
Dual sim	0	1	0.51	0.5
5G	0	1	0.52	0.5
4G	0	1	0.52	0.5

Class is Price class to predict whether the mobile is very economical, Economical, Expensive or very Expensive. Basically price is also a continuously changing real value, but it is mapped into above four classes with following criteria.

Table 2. Classification criteria

Price Range	Class
0	Very economical
1	Economical
2	Expensive
3	Very Expensive

So a regression problem is converted into classification. Because the main weakness of decision trees and naive bayes classifier is their inability to handle output classes with numeric values. Hence, the price attribute had to be classified into classes which contained a range of prices but this evidently introduced further grounds for inaccuracies.

To evaluate the classifier performance data is into Training set and test set.

3.2 Exploratory Data Analysis (EDA)

In this section, we have taken different functions to analyse the data. The functions are : describe(), info(), shape, etc. We have also plotted the graphs to show the relation among different features of the dataset. After this we have used the drop() function to remove the features from test and train dataset which are not required. This process is known as dimensionality reduction.

Dimensionality reduction is the process of reducing the number of random variables (Features) under consideration, by obtaining a set of principal variables. The higher the number of features, the harder it gets to visualize the training set and then work on it. Sometimes, most of these features are correlated, and hence redundant. This is where dimensionality reduction algorithms come into play. Two types of Dimensionality reduction algorithms are there i.e. Feature selection, Feature extraction.

After performing dimensionality reduction, we split the train data into X_train, Y_train, X_test and Y_test. The splitting of train data helps in finding the metrics or we can say the accuracy score. Then we have performed the normalisation or scaling of the dataset, which is required for the algorithms like KNN, random forest classifier etc.

3.3 Classification

For training the model we have used the different classification models in order to have the good accuracy or better performance of the model.

The algorithms used are : Decision tree, K nearest neighbors(KNN), Random forest classifier, Logistic regression, Support vector machine(SVM). So, these classification algorithms helps in prediction of price.

4. EVALUATION

Now let's go through the last step i.e. evaluation. The major aim of this project is to predict price range of the smart mobile phones. We have imported different packages which are used in this project. The packages used are pandas, matplotlib, scikit-learn(sklearn). Pandas mainly works with the relational and labelled data both easily and intuitively. It provides various data structures and operations for manipulating numerical data and time series. Matplotlib is a multiplatform data visualization library which is used to plot the graphs. It is used for creating static, animated and interactive visualizations. The scikit-learn provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python.

There are various algorithms in Machine learning, so choosing the best algorithm for the given dataset and problem is the main point to remember while creating a machine learning model.For training the machine learning model we have used different classification algorithms like Decision tree, K nearest neighbors(KNN), Random forest classifier, Logistic regression, Support vector machine(SVM).

We have used the metric of accuracy for the evaluation of the models. Accuracy is one of the simplest metrics available to us for classification models. It is the number of correct predictions as a percentage of the number of observations in the dataset. This metric accurately assess the performance of the models used for training.

Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. The accuracy score obtained using decision tree algorithm is 0.86.

K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm. The accuracy score obtained using this algorithm is 0.5075.

Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1. It gives the best accuracy score i.e. 0.96.

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. The accuracy score obtained is 0.8725.

SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. The accuracy score obtained using svm is 0.9475.

5. COMPARATIVE STUDY

Comparison in machine learning is done in terms of maximum accuracy and minimum number of features selected. Maximum accuracy means more data classified correctly. While minimum number of feature means minimum memory required and reduced computation complexity. We have performed the comparative study of different algorithms used during the evaluation of metric.

The comparison is made by plotting the bar graph between algorithms and their accuracy score. With the help of evaluation we get to know about the accuracy score of the algorithms used for training the model. For each algorithm, we get the different accuracy score.

Table 3. Accuracy Score

Algorithms	Accuracy Score
Decision Tree	0.86
KNN	0.5075
Random Forest Classifier	0.8725
Logistic Regression	0.96
SVM	0.9475

In the above table we can see that the Logistic regression is having the best accuracy score while the KNN algorithm is having the minimum accuracy. So, it shows that the logistic regression algorithm is predicting the prices more accurately. After logistic regression the other algorithms like SVM, random forest classifier and decision tree are also giving the good results for predicting the price range of smart mobile phones.

6. CONCLUSION

This work can be concluded with the comparable results of algorithms and classifiers like Decision tree, random forest classifier, logistic regression and support vector machine except the KNN, as it gives the minimum accuracy even by removing the extra features or by performing dimensionality reduction . The combination of Logistic regression and the accuracy score metric has achieved maximum accuracy and selected minimum but most appropriate features. The main reason of low accuracy rate in KNN is low number of instances in the data set. The metric which is yielding the good accuracy is the accuracy score metric while the other metrics like mean squared error, absolute error and f1 score, are having the lowest accuracy for the given dataset. So, for predicting the price of smart mobile phones, the logistic regression comes up with the best accuracy and predict the price range more accurately.

7. OUTCOMES OF THE WORK

- Cost prediction is the very important factor of marketing and business. To predict the cost same procedure can be performed for all types of products for example Cars, Foods, Medicine, Laptops etc.
- Best marketing strategy is to find optimal product (with minimum cost and maximum specifications). So products can be compared in terms of their specifications, cost, manufacturing company etc.
- By specifying economic range a good product can be suggested to a costumer.

8. FUTURE WORK EXTENSION

- More sophisticated artificial intelligence techniques can be used to maximized the accuracy and predict the accurate price of the products.
- Software or Mobile app can be developed that will predict the market price of any new launched product.
- To achieve maximum accuracy and predict more accurate, more and more instances should be added to the data set. And selecting more appropriate features can also increase the accuracy. So data set should be large and more appropriate features should be selected to achieve higher accuracy.

9. REFERENCES

[1] https://www.kaggle.com/code/vikramb/mobile-price-prediction

[2] Learn Mobile Price Prediction Through Four Classification Algorithms (analyticsvidhya.com)

[3] Shonda Kuiper, "Introduction to Multiple Regression: How Much Is Your Car Worth? " , Journal of Statistics Education · November 2008

[4] Mariana Listiani , 2009. "Support Vector Regression Analysis for Price Prediction in a Car Leasing Application". Master Thesis. Hamburg University of Technology.

[5] Limsombunchai, V. 2004. "House Price Prediction: Hedonic Price Model vs. Artificial Neural Network", New Zealand Agricultural and Resource Economics Society Conference, New Zealand, pp. 25-26. 2004

[6] Sameerchand Pudaruth . "Predicting the Price of Used Cars using Machine Learning Techniques", International Journal of Information & Computation Technology. ISSN 0974-2239 Volume 4, Number 7 (2014), pp. 753- 764