Prediction of Phone Prices Using Machine Learning Techniques

S. Subhiksha, Swathi Thota and J. Sangeetha

Abstract In this modern era, smartphones are an integral part of the lives of human beings. When a smartphone is purchased, many factors like the display, processor, memory, camera, thickness, battery, connectivity and others are taken into account. One factor that people do not consider is whether the product is worth the cost. As there are no resources to cross-validate the price, people fail in taking the correct decision. This paper looks to solve the problem by taking the historical data pertaining to the key features of smartphones along with its cost and develop a model that will predict the approximate price of the new smartphone with a reasonable accuracy. The data set [12] used for this purpose has taken into consideration 21 different parameters for predicting the price of the phone. Random forest classifier, support vector machine and logistic regression have been used primarily. Based on the accuracy, the appropriate algorithm has been used to predict the prices of the smartphone. This not only helps the customers decide the right phone to purchase, it also helps the owners decide what should be the appropriate pricing of the phone for the features that they offer. This idea of predicting the price will help the people make informed choice when they are purchasing a phone in the future. Among the three classifiers chosen, logistic regression and support vector machine had the highest accuracy of 81%. Further, logistic regression was used to predict the prices of the phone.

Keywords Support vector machine · Logistics regression · Smartphone prices · Random forest

S. Subhiksha (✉) · S. Thota · J. Sangeetha
School of Computing, SASTRA Deemed University, Tirumalaisamudram, Thanjavur, Tamil Nadu 613401, India
e-mail: subhisru@gmail.com

K. S. Raju et al. (eds.), *Data Engineering and Communication Technology*,
Advances in Intelligent Systems and Computing 1079,
https://doi.org/10.1007/978-981-15-1097-7_65

1 Introduction

On a daily basis, we encounter a lot of trade-off in life, for instance, what to opt tasty food versus healthy food, cost of product versus features of product, durability versus reliability and lot more. The complexity of such situations increases day by day and common man faces a lot of tough time to cope up with it. Moreover, taking correct decision in limited time has always been crucial in such a scenario. Nowadays, in this modern digital age, social media is substantially evolving at a very fast pace. The growth has always been tremendous by connecting everyone across the globe in a quick, secure and convenient manner. Smartphones are one of the most readily accessible devices by every individual in this platform. It is one of the most common devices which many individuals possess and it is quite impossible for anyone to survive without it. With various customizations, features along with add-on plugin have enhanced its position in market.

The increase in demand for smartphones has simultaneously led to increase in manufacturers all over the world. The manufactures have started increasing features of their product to securely speed up their position in market. This arises problem for people as to which smartphone to purchase with appropriate features. Overall, purchase of smartphone has always been an issue encountered by all in some instance of time. People spent a lot of time thinking and cross-checking with their peers about product. People are often in dilemma whether the features provided by the manufacturer of the phone are really worth the cost of buying. The attempts to purchase a phone by people in dilemma led to disappointing results of spent significant amount of time as well as their money. "These above-mentioned factors fascinate up the thought of 'Predicting the price of Mobile Phone"'. This helps people in making correct decision as well as for manufacturer for validating cost of phone with features provided to their customers. At the end of day, both customer and manufactures get satisfied with product on the whole with valid statistical proofs.

In social media, many people tend to post rating of product without any hesitation. So, in this research, the historical sentiments of people are taken into consideration, i.e. their opinion of price whether it is high, medium or low. Apart from the important factors like display, processor and memory, other features like camera, thickness, battery and connectivity have also been taken into account to determine the approximate price for a phone.

For predicting the price, we are using data set [1] from the popular data set platform Kaggle. The train and test data set are given as two different CSV files. So, to train the model and for predicting which model gives the most accuracy, the train data set is used. The trained model is then used on the test data set to predict the price.

For this concept, due to unavailability of related specific papers, papers regarding recommender systems, sentiment analysis and other predicting system have been used for the purpose of literature survey.

In the paper [2], the authors have mined for historical data from many sources like IMDB and Rotten Tomatoes. After processing the data, the authors have implemented SVM and neural networks to find the accuracy of the prediction. For the data set chosen by the team, the neural network gave the most accurate predictions.

Mansouri et al. in the paper [3], have made prediction on the useful battery charge that can be used in a UAV. The authors have used a variation of SVM, an advanced tree-based algorithm, a linear sparse model and a multilayer perceptron to make the prediction. Using all these algorithms, a preliminary investigation was done.

In the paper [4], the authors look to predict the performance of Karachi Stock Exchange. Various machine learning techniques like single-layer perceptron, SVM, radial basis function and multilayer perceptron are used on the data set to make the prediction. The performance of the multilayer perceptron was the best. The final conclusion drawn from the research was that KSE performance can be predicted using machine learning techniques.

In the paper [5], neural network is implemented and an accuracy of 75% is achieved. This sentiment analysis is very useful for mining useful knowledge from all the data available.

Paper [6] aimed to recommend users the ingredients for different cuisines. The recipes were checked using classifiers like support vector machine and associative classification. The accuracy used in classifiers was used for comparison.

The paper [7] predicts factors that will be affecting future usage of tags. Tags are essentially used to easily search for the answer in that particular domain. Machine learning techniques are used to automate as well as classify them popularity of tags based on structural and non-structural features. The classifiers used were logistic regression, SVM, random forest and AdaBoost. Random forest is the best among the other classifiers based on accuracy.

The paper [8] helps in proper usage of correct algorithm in recommended system and identifying machine learning techniques, new research areas to invest upon. It also focuses on main and alternative performance metrics.

The paper [9] attempted to research efficient models and compared their performance in predicting the direction of movement of daily Istanbul stock Exchange (ISE) National 100 index. Artificial neural network and support vector machine were used for classification. Among them, artificial neural network model (ANN) is a better performance model compared to support vector machine (SVM).

The paper [10] helps in generating forecast of online content. Prediction methods of existing system are used in algorithm for real-time forecasting of popularity with minimal training information.

The paper [11] helps in predicting severity of crash that is expected to occur in future. It uses multinomial logit (MNL), nearest neighbor classification (NNC), support vector machines and random forest (RF) for predicting the traffic crash severity. The proposed approach showed NNC as the best prediction performance in overall and is in more accurate in severe crashes.

Based on the literature, an efficient and accurate way for predicting the price of a smartphone has been determined. These are discussed in the corresponding sections of the paper. Section 2 of the paper talks about the various classifiers used. Sections 3 and 4 discuss about the various classifiers used along with the proposed work. Section 5 performs accuracy analysis of the result obtained. Finally, the paper is concluded in Sect. 6.

2 Classifiers

2.1 Random Forest Classifier

Random Forest Classifier is one of the decision tree algorithms. As the name forest suggests, the name forest refers to the fact that the algorithm is a collection of many decision trees. When a number of cases are present in the training set, the equivalent number of samples is taken at random and is used for training the growing tree. Each of the decision tree is let to grow for the maximum extent. Pruning is not done [12, 13].

Random forest is an aggregate of decision trees, supervised classifier. Each decision tree is in turn a subset of forest which is trained on different training data set with some random selected subset of features. The final result is collection of results of different subset model.

2.2 Support Vector Machine

Among the various classification algorithms available, this is a very powerful algorithm. Here, the data are first plotted as a scatter plot. Then, multiple lines are drawn separating the various data points. Then, a line is modeled in such a way that it splits the data points into two distinct groups [12].

Support vector machine is primarily used for modeling binary classifiers, which classifies data into two categories by developing hyper plane in multidimensional space. Moreover, its decision of predicting is based on linear function [14].

$$y = f(x) = a + bx \tag{1}$$

It is supervised model wherein it learns from training data and predicts the output. The best hyper plane is known by name called margin hyper plane which maximizes sum of distance on either.

The Rbf parameter in support vector machine means radial basis function. It is widely used in the algorithms that use kernels for execution. One such algorithm that uses Rbf is the SVM algorithm. The equation representing the equation is given below [15].

$$\exp\left(-\frac{1}{2}|x-x'|^2\right) = \sum_{j=0}^{\infty}\frac{1}{j!}\left(x^T x'\right)\exp\left|\frac{-1}{2}\right|x^2|\exp|\frac{-1}{2}|x^2|\tag{2}$$

The above equation shows the equation of radial basis function kernel.

2.3 *Logistic Regression*

The name regression for this algorithm is a misnomer. This is a classification algorithm. By using a given set of independent variables, this algorithm predicts discrete values as output. Mathematically, the logarithm of probability of outcome is modeled as a combination of various predictor variables. This combination is linear in nature. Here, the algorithm chooses parameters in such a way that odds of observing the sample quantity is increased. Mathematical equation is as follows [16]:

Linear Equation

$$y = a + bx$$

$$\text{Outcome probability} = \frac{P}{(1-P)}$$

$$\text{loge}\left(\frac{p}{(1-p)}\right) = a_0 + a_1x_1 + a_2x_2 + \cdots + a_n x_n$$

The function is a step function, and so, logarithmic value is taken to make replication easier [12].

3 Data Description and Methodology

The data set [1] was downloaded from Kaggle. As the preprocessed data were already available no preprocessing has been done on the data set. The data set has the following columns. Each of this is an attribute associated with phone. The test data set that has the same attributes as the train dataset except the price column is not available. This data set takes into account all the parameters associated with a smartphone as they are all a deciding factor on the price of the phone. Listed below is the various names of the column in the table.

- Battery_power: Total energy that a battery can store at a time (measured in mAh)
- Blue: if phone as Bluetooth or not

- Clock_speed: speed at which the microprocessor in the phone executes the instructions
- Dual_sim: If dual-sim support is offered or not
- Fc: Front camera focus in megapixels
- Four_g: If 4G is available or not
- Int_memory: Internal memory of phone in gigabytes
- M_dep: Mobile depth in cm
- Mobile_wt: Weight of the phone
- N_cores: Number of cores in the processor
- Pc: Primary camera focus in megapixels
- Px_height: Pixel resolution height
- Px_width: Pixel resolution width
- Ram: Random access memory (megabytes)
- Sc_h: Screen height of phone in cm
- Sc_w: Screen width of phone in cm
- Talk_time: maximum time that a single battery charge will last when used for talking
- Three_g: Has 3G or not
- Touch_screen: is the display touch screen or not
- Wi-Fi: Has Wi-Fi connectivity or not
- Price_range: The target variable of this data set. This has four discrete values of 0(low cost), 1(medium cost), 2(high cost) and 3(very high cost).

Here, the approximate price range is predicted. We use discrete value for the price in the place of exact price. The price of the column is predicted based on 21 different parameters. The data set contains data of close to 2000 odd phones.

4 Proposed Work

The technique proposed in this paper is to take different features of phone affecting the price in market as the input of problem instance. The output of model predicts whether the price of phone is high, medium or low. Prior to classification of model, data preprocessing task is to be performed on the data set [1]. But the data set considered here was already preprocessed and was ready for use. The supervised classifiers like logistic regression, random forest and support vector machine were used for training the data set. The various parameters associated with the classifiers were modified so as to make sure that the make sure that the model was not overstrained. The accuracy of the trained model was observed for all the supervised algorithms used. Logistic regression algorithm was found to be giving the most accurate result of 81% alongside SVM which also equivalently gave the same result.

After this training based on the accuracy, the price range of the phone was predicted using the logistic regression model. The price range was returned as an

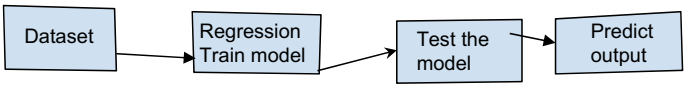

Fig. 1 A diagram representing the flow of events in the proposed model of execution of this project

array. All the coding for the implementation was done using Python language. Various libraries like pandas, sklearn were used to help visualize the data so that the correct model could be predicted for use. Implementation of the supervised algorithm on the data set [1] gave the following accuracies for the different algorithms implemented. Figure 1 represents the accuracy portrayed by the various classifiers along with the changes made in their parameter used during the implementation has been drawn.

- **Support Vector Machine**

  SVM Linear Accuracy: 0.81
  SVM Rbf Accuracy: 0.77

- **Logistic Regression**

  Linear model accuracy: 0.81

- **Random Forest**

  Accuracy: 0.77
  Best max depth: 13 (Fig. 2).

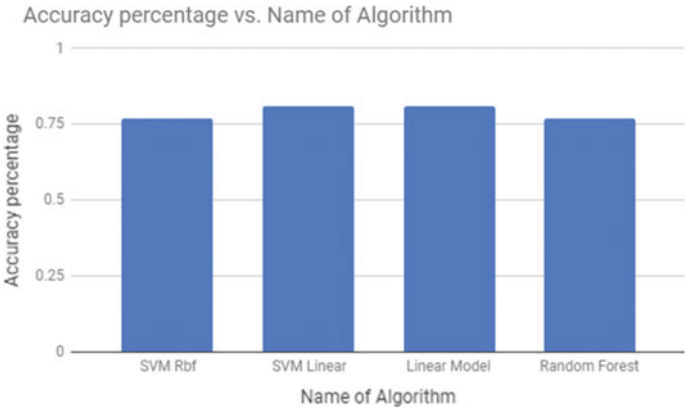

Fig. 2 A graph representing the accuracy percentage as a measure against the algorithm for which the accuracy was obtained. Even the accuracy that was obtained upon changing the parameter for the algorithms is represented here

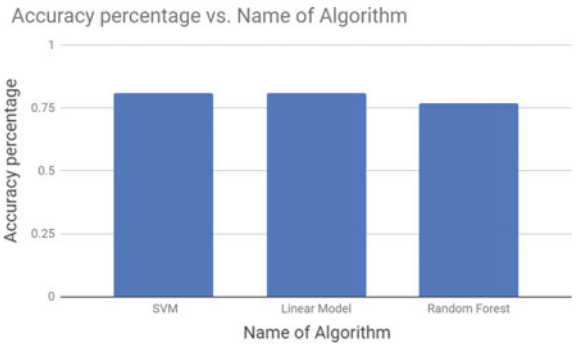

Fig. 3 A graph representing the accuracy percentage as a measure against the algorithm for which the accuracy was obtained

5 Performance Measures and Analysis

Accuracy is the measure of the correctness of result when a test is performed upon a set of values. In machine learning, accuracy refers to the correctness with which a trained model is able to predict the output value in the test data set. First accuracy is measured so that we can choose a data set appropriate algorithm and train the algorithm to our needs to give the most accurate prediction. For the same reasons, on the data set chosen here, three different algorithms were used.

Based on the literature reading performed on the various papers, the various algorithms chosen to predict the accuracy of the trained model were logistic regression, support vector machine and random forest.

For logistic regression, a variation in parameter was done. The accuracy obtained was the same when the inverse of regularization parameter was used. The accuracy obtained was 81%.

In random forest classifier, the primary accuracy was found for all the depths. From the result obtained, branch with the most accuracy was chosen along with its predecessor and successor. These branches were further expanded to find the branch which gives the best accuracy. When this classifier was used on the train data set, the maximum accuracy obtained was 77%.

For support vector machine, two different variations out of two were implemented on the data set. The SVM linear algorithm gave an accuracy of 81%. The algorithm of SVM with the Rbf parameter on the other hand gave an accuracy of only 77%. In Fig. 3, the maximum accuracy of the various classifiers has been represented graphically.

6 Conclusion and Future Work

The logistic regression algorithm and SVM were found to give the most accuracy of 81%. Then, the prediction of the price is done using logistic regression. This is just a preliminary paper. The future work to be done on this paper lies in predicting the

approximate price of a second-hand phone sale. The same project can further be extended to predict the exact price of the phone instead of the discrete values that is being predicted in this paper. By further pursuing this project, we will be able to help people to spend money wisely and also mine other data like the how the price is affected by the brand, how long a phone can work, etc.

References

1. https://www.kaggle.com/iabhishekofficial/mobile-price-classification#train.csv
2. Quader, N., Ganim M.O., Chaki, D., Ali, M.H.: A machine learning approach to predict movie box-office success. In: 2017 20th International Conference of Computer and Information Technology (ICCIT), pp. 1–7. IEEE (2017)
3. Mansouri, S.S., Karvelis, P., Georgoulas, G., Nikolakopoulos, G.: Remaining useful battery life prediction for UAVs based on machine learning. IFAC-Papers OnLine **50**(1), 4727–4732 (2017)
4. Usmani, M., Adil, S.H., Raza, K., Ali, S.S.: Stock market prediction using machine learning techniques. In: 2016 3rd International Conference on Computer and Information Sciences (ICCOINS), pp. 322–327. IEEE (2016)
5. Ramadhani, A.M., Goo, H.S.: Twitter sentiment analysis using deep learning methods. In: 7th International Annual Engineering Seminar (InAES) pp. 1–4. IEEE (2017)
6. Jayaraman, S., Choudhury, T., Kumar, P.: Analysis of classification models based on cuisine prediction using machine learning. In: International Conference on Smart Technologies For Smart Nation (SmartTechCon) pp. 1485–1490. IEEE (2017)
7. Fu, C., Zheng, Y., Li, S., Xuan, Q., Ruan, Z.: Predicting the popularity of tags in StackExchange QA communities. In: 2017 International Workshop on 2017 Complex Systems and Networks (IWCSN) pp. 90–95. IEEE (2017)
8. Portugal, I., Alencar, P., Cowan, D.: The use of machine learning algorithms in recommender systems: a systematic review. Expert Syst. Appl. **1**(97), 205–227 (2018)
9. Kara, Y., Boyacioglu, M.A., Baykan, Ö.K.: Predicting direction of stock price index movement using artificial neural networks and support vector machines: the sample of the Istanbul Stock Exchange. Expert Syst. Appl. **38**(5), 5311–5319 (2011)
10. Lymperopoulos, I.N.: Predicting the popularity growth of online content: model and algorithm. Inf. Sci. **10**(369), 585–613 (2016)
11. Iranitalab, A., Khattak, A.: Comparison of four statistical and machine learning methods for crash severity prediction. Accid. Anal. Prev. **30**(108), 27–36 (2017)
12. https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/
13. https://thedatalass.com/2018/04/17/random-forest/
14. https://www.researchgate.net/figure/Figure-3-SVM-classification-scheme-H-is-the-classification-hyperplane-W-is-the-normal_fig3_286268965
15. https://en.wikipedia.org/wiki/Radial_basis_function_kernel
16. https://www.saedsayad.com/logistic_regression.htm