MOBILE PRICE PREDICTION USING WEKA

[1]Pritish Arora, [2]Sudhanshu Srivastava, [3]Bindu Garg

[1]UG Student, [2]UG Student, [3]Professor
Department of Computer Engineering,
Bharati Vidyapeeth (Deemed to be University) College of Engineering, Pune India

***Abstract*: The key purpose of this research work is to determine "If the mobile with given features would be under a certain price range." Specific feature selection algorithms are used to recognize and delete features that are less necessary and redundant, and have minimal complexity in computation. Different classifiers are used to achieve the best possible accuracy. Results are measured in terms of achieving the maximum accuracy and choosing the minimum features. Statement is made based on the algorithm for best selection of features and best classifier for the given dataset. This work can be used to find the optimal product (with minimum cost and maximum features) in any form of marketing and industry. It is suggested that future work will extend this research and find a more sophisticated solution to the given problem and a more accurate tool for estimating prices.**

***Keywords*: Machine Learning, Prediction, Decision Tree, Naïve Bayes**

1. INTRODUCTION

Price is the marketing and business attribute which is the most powerful. The costumer's very first question is about the price of the things.

All the customers are worried first and wonder "whether he can buy something with the requirements given or not."So the basic aim of the research is to estimate price at home. This paper represents only the first step towards the destination described above.

Artificial intelligence — which makes the computer intelligently capable of answering the questions — is now a very large field of engineering. Machine learning provides us with the latest artificial intelligence methods, such as classification, regression, supervised learning and unsupervised learning and much more. Different machine learning tools are available, such as MATLAB, Python, cygwin, WEKA etc. We may use one of Decision tree, Naïve Bayes and several other classifiers. Different types of algorithms are necessary for selecting only the best features and reducing the dataset. That will lower the problem's computational complexity. Because this is the problem of optimization, several optimization techniques are often used to reduce the dataset's dimensionality.

Mobile now a day is one of the apps with the most sales and purchases. New mobiles are released each day with new version and more apps. Hundreds and thousands of cell phones are sold and bought every day. So here the prediction of the mobile price class is a case study for the given type of problem i.e. finding an optimal product. The same work can be done to estimate the real price of all goods, such as vehicles, motorcycles, generators, engines, food items, medication, etc.

Several apps are very important for estimating mobile prices, for example Mobile Processor. In today's busy human life the timing of batteries is also very critical. Mobile size and thickness are also important determinants of decision. Internal memory, camera pixels and the consistency of the video must be remembered. Internet surfing is also one of the most significant limitations of this 21st-century technological period. And so is the list of several features dependent on those, it is decided on mobile size. So we're going to use all of the above features to determine whether the smart-phone will be very-economic, economical, expensive or very-costly.

Paper structure is as follows. Next section is analysis of previous work.3 rd Section involves Technique and Experimental procedure. Section 4 contains a description of the results. Comparative analysis is conducted in section 5. After the paper in section 6 is finalized. The results of the work are dealt with in section 7. Some suggestions about future research are finally given in 8th section.

2. PREVIOUS WORK

An interesting research background for machine-learning researchers is the use of previous data to predict the price of available and new launch product. Prices of second-hand cars in Mauritius are estimated by Sameerchand-Pudaruth[1]. To predict the prices, he introduced many techniques such as multiple linear regressions, k-nearest neighbors (KNN), Decision Tree and Naïve Bayes. From all these techniques Sameerchand-Pudaruth got comparable results. It was found during research that most common algorithms, i.e. Decision Tree and Naïve Bayes, are incapable of handling, classifying and predicting numeric values. There were only 97 (47 Toyota+38 Nissan+12 Honda) examples for his work. Very poor prediction accuracies were recorded because of lesser number of instances used [1].

Shonda Kuiper [2] worked in the same field too. To estimate the price of 2005 General Motor cars, Kuiper used multivariate regression model. He collected the data from www.pakwheels.com, available online source. Main part of this research work is "Introduction of suitable techniques for variable selection, which helped to find out which variables are more suitable and relevant for model inclusion. This (His research) allows students and prospective researchers in many fields to consider the conditions under which studies are to be performed and gives them the ability to determine when correct techniques should be used[2].

The definition of supporting vector machine (SVM) is used for the same work by one other researcher, Mariana Listiani[3]. Listiani used the above mentioned technique to forecast prices of leased vehicles. In this study, it has been found that the SVM methodology is much better and more reliable for price prediction compared to other such as multiple linear regressions when there is a very large data set. The researcher also showed that SVM is also effective at managing high-dimensional data and preventing both the under-fitting and over-fitting problems. Genetic Algorithm used to identify essential features for SVM Listiani. The technique however failed to demonstrate why SVM is better than simple multiple regression in terms of variance and mean standard deviation [3].

Neural Networks (NN) are excellent at estimating house prices, this has been concluded in the Limsombunchai research [4]. His method was more reliable when compared with hedonic method. All methods work the same except the model is first trained in NN and then evaluated for prediction. Using both methods NN generated higher R-sq and lower root mean square error (RMSE), whereas lower values were provided by the hedonic. This study was limited because the real house price was lacking and the analysis work only used estimated prices[4].

K Noor and Saddaqat J[5] have experimented with various techniques to estimate the price of the vehicles. Using multiple linear regressions the researchers achieved maximum accuracy. This paper proposes a system where price is predicted based on variable, and that price is derived from factors such as vehicle type, make, area, edition, colour, mileage, alloy rims and power steering [5].

3. METHODOLOGY

The experiment is carried out using WEKA (Waikato Setting for the Study of Knowledge). The key machine-learning steps are as follows

3.1 Data Collection

Ten smartphone apps are collected from www. kaggle.com[6] i.e.

Catagory(whether Apple, Samsung, Lenovo, NOKIA etc make the specified smartphone). Memory card slot is regarded as functionality whether or not it is present.

The monitor size(Inches), the weight(g), the thickness(mm), the internal memory size(GB), the camera pixels(MP), the video quality, the RAM size(GB) and the battery (mAh) all have real values with the following distinctions.

Class is Price class for determining whether the mobile is Very economic, Economical, Competitive or Very Competitive. Basically price is still a continuously evolving real value, but with the following criteria it is divided into over four classes.

Thus the problem of regression is converted into classification. Because the major weakness of decision trees and naive bayes classifier is their inability to handle numeric values for output classes. The price attribute therefore had to be divided into classes comprising a variety of prices, but this naturally gave rise to additional explanations for inaccuracies [1].

The output data of the classifier is split into training set and test set, 108 training instances and 28 test set instances (total of 134 instances).

3.2 Dimensionality Reduction

Reduction of dimensionality is the method of the number of random variables (Features) under consideration, by obtaining a collection of key variables [7]. The higher the number of features, the more difficult it is for the training set to be visualized and then worked on. Most of these features often are linked, and thus redundant. It is here that algorithms of dimensionality reduction come into play [7].

There are two types of algorithms for Dimensionality reduction ie selection of features, extraction of features.

3.2.1 Feature Selection

In the selection of features we are interested in finding k of the d dimensions that give us the most detail, and we discard the other dimensions $(d - k)$ [8].

3.2.2 Feature Extraction

We are interested in seeking a new set of k dimensions in the extraction of features that are variations of the original d dimensions, for example Principal Component Analysis [8].

Algorithms for selection of the apps are used here. There are two approaches: choosing forwards and choosing backwards.

3.2.1.1 Forward Selection

In forward selection, we start with no variables and add them one by one, adding the one at each stage that most decreases the error, until any addition does not decrease the error (or only slightly decreases it).

3.2.1.2 Backward Selection

In reverse selection we start with all variables and remove them one by one, removing at each step the one that most decreases the error (or only slightly increases it), before any further removal significantly increases the error [8].

InfoGainEval and WrapperattributEval are used on two function selection algorithms. InfoGainAttributeEval measures the value of an attribute by calculating the benefit in knowledge about the class [9]. It brings us the listed list from the most important feature to the least important features.

Thus WrapperattributeEval is thought to "tie" the process of extraction of features around the learner that it uses as a subroutine [8]. It just offers us a rundown of essential features.

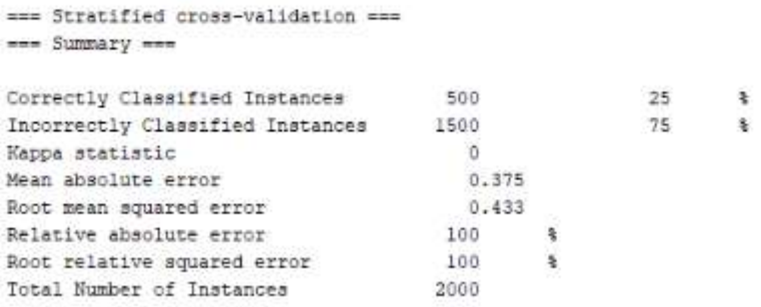
3.3 Classification

Now let's go through the final stage which is classification. As mentioned above, separate test set is used for classifier evaluation and accuracy finding. Any classification is accurate if it can be determined by measuring the number of class samples correctly identified (true positives), the number of samples correctly identified which are not class members (true negative) And samples which were either misallocated to the class (false positives) or not marked as class samples (false negatives)[10]. Accuracy tells us percentage of instances that are categorized correctly. In mathematics

Accuracy □ (Correctly Classified Samples / Total Samples) * 100

4. RESULTS

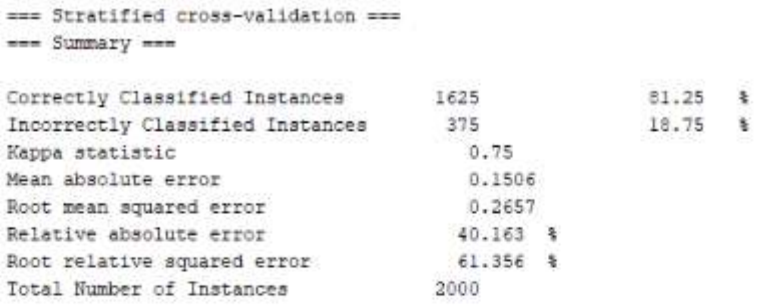
The result here is the accuracy of the machine learning model which is accessed by a comparative study of the accuracy of various machine learning models. Summary of various models is provided as follows-



```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         500               25      %
Incorrectly Classified Instances      1500               75      %
Kappa statistic                          0
Mean absolute error                      0.375
Root mean squared error                  0.433
Relative absolute error                100      %
Root relative squared error            100      %
Total Number of Instances             2000
```

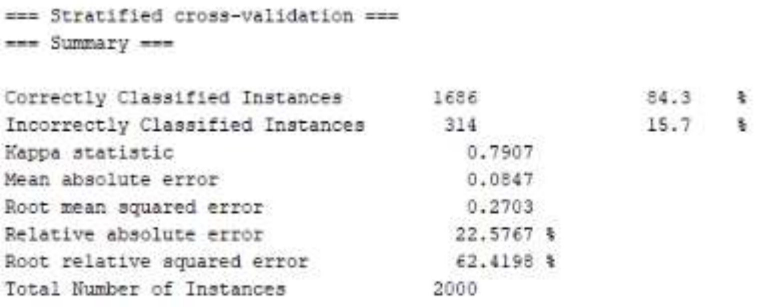
Fig 1: Classification results for ZeroR algorithm in WEKA



```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1625               81.25   %
Incorrectly Classified Instances       375               18.75   %
Kappa statistic                          0.75
Mean absolute error                      0.1506
Root mean squared error                  0.2657
Relative absolute error                 40.163  %
Root relative squared error             61.356  %
Total Number of Instances             2000
```

Fig 2: Classification results for Naive Bayes algorithm in WEKA



```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1686               84.3    %
Incorrectly Classified Instances       314               15.7    %
Kappa statistic                          0.7907
Mean absolute error                      0.0847
Root mean squared error                  0.2703
Relative absolute error                 22.5767 %
Root relative squared error             62.4198 %
Total Number of Instances             2000
```

Fig 3: Classification results for J48 decision tree in WEKA

Comparing the results, the maximum accuracy is obtained by the J48 Decision tree. All features are selected for this specific dataset as a more generic and the most optimal model is to be devised which gives maximum accuracy and hence can be later configured according to the specificity of another dataset of the similar domain.

5. CONCLUSION

Cost estimation is the marketing and business aspect which is very significant. For all types of goods, for example vehicles, foods, medication, laptops etc., the same technique can be done to predict the cost. The best marketing strategy is to find the optimum product (with minimum cost and maximum specifications). Products can thus be measured in terms of their requirements, prices, Production Company etc. A good product can be recommended to a costumer by defining the economic range which can best be achieved by mining and analysis of data. In our use case the price range of a mobile was successfully predicted with a high accuracy by training the model for a dataset of two thousand instances with various attributes using J48 decision tree learning model in WEKA tool.

REFERENCES

[1] Pudaruth Sameerchand. "Predicting the price of used vehicles using machine learning techniques," International Information and Computer Technology Magazine.
ISSN 0974-2239 Volume 4, Number 7 (2014), pp. 753- 764
[2] Shonda Kuiper, "Multiple Regression Introduction: How much is your car worth? "November 2008, Journal of Statistics Education

[3] Mariana Listiani , 2009. “Support Vector RegressionAnalysis for Price Prediction in a Car LeasingApplication”. Master Thesis. Hamburg University of Technology.
[4] 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
[5] Kanwal Noor and Sadaqat Jan, “Vehicle Price Prediction System using Machine Learning Techniques” , International Journal of Computer Applications (0975 – 8887) Volume 167 – No.9, June 2017.
[6] Mobile data and specifications online available from https://www.gsmarena.com/ (Last Accessed on Friday, December 22, 2017, 6:14:54 PM)
[7] Introduction to dimensionality reduction, A computer science portal for Geeks.https://www.geeksforgeeks.org/dimensionality-reduction/ (Last Accessed on Monday , Jan 201822, 3 PM)
[8] Ethem Alpaydın, 2004. Introduction to Machine Learning, Third Edition. The MIT PressCambridge, Massachusetts London, England
[9] InfoGainAttributeEval-Weka Online available fromhttp://weka.WrapperattributEval/doc.dev/weka/attributeSelection/InfoGainAttributeEval.html (LastAccessed in Jan 2018)
[10] Thu Zar Phyu, Nyein Nyein Oo. Performance Comparison of Feature Selection Methods. MATEC Web of Conferences42, (2016).