Mobile Price Prediction Using ML Algorithms

Gautam Dhall[1,] *Dr.S.Nithiya*[2]
[1]*Btech Scholar , Department Of Computer Science Engineering , SRM Institute OF Science And Technology , Kattankulathur , Chennai(603203) , India*
2Assistant Professor,Department Of Computer Science Engineering , SRM Institute OF Science And Technology , Kattankulathur , Chennai(603203) , India

ABSTRACT

The key cause of this study paintings is to decide whether a given cell phone with given capabilities could fall under which positive rate range. Various characteristic picks algorithms are used to understand and delete diverse capabilities from the information set which are much less important and redundant and feature a completely minimum complexity in the computation. Many Different algorithims are being used in such to attain the exceptional feasible accuracies. Results are particularly measured in phrases of attaining the most of the accuracies and selecting the minimal capabilities. The statements are made primarily based on the set of rules for exceptional capabilities and exceptional classifiers for the given dataset. These paintings are used to discover the premiere product (with minimal price and most capabilities) in any shape of advertising and marketing and industry. It is usually recommended that destiny paintings will make bigger this studies and discover a greater state-of-the-art method to the given trouble and a greater correct device for estimating prices.

Keyword :- *Feature Engineering, SVM, Naive Bayes,Machine Learning*

1.Introduction

Price is an advertising and enterprise characteristic that is the maximum power. The consumer's first actual query is ready the charge of all things. All clients are involved first and they wonder if they should get something according to them with the necessities given. Hence, the simple purpose of this study is to predict the charge of cellular phones. This paper presents the handiest step toward the vacation spot defined above. AI — "which makes the pc intelligently able to answer the questions" is now a complete massive area of engineering technology. Machine studying presents us with the modern synthetic methods, which includes , regression, classification, supervised studying and unsupervised studying and lots greater. Different system studying gear are available. One such is by using Python. We also use certainly considered one among SVM, Decision Tree , XgBoost and numerous classifiers. Different algorithms are important for choosing satisfactory functions and lowering the dataset. That will decrease the trouble caused by computational complexities. Because of it leads to the trouble of the optimizations. Numerous optimization strategies are frequently used by us to lessen the dataset dimensionality.

Mobile apps are one of the many things with the maximum income and daily purchases. Many new mobile phones are launched every day with a new edition and greater apps. Thousands of various cell phones are offered and are purchased each day. Hence, the estimation of the cell charge magnificence is observed for the given kind of trouble i.e. locating a premiere products. The identical paintings may be performed to predict the actual charge of all goods. Several applications are very critical for estimating cell prices, for an instance Mobile processors. In modern busy humans existence, the battery are likewise very critical. Mobile length also is critical determinants of decision. Memory, digital digicam pixels ought to be remembered. On Internet, browsing is likewise one of the maximum large boundaries of this "21st century"technology period. And hence, the listing of numerous functions depending on those, it's miles determined on cell length. So we are going to consider the above functions to decide whether the smartphone can be very affordable, affordable, highly -expensive & very expensive to afford.

2. Methodology

2.1 Data Collections

10 phone applications are accumulated from Kaggle.com[1] Memory card slot appears as capability whether or now no longer it's far present.

The reveal size, thickness,weight, the inner reminiscence size, the digital digicam pixels(MP), the battery all have actual values with the subsequent distinctions.

Class is a price elegance for figuring out if the cellular is very cheap, affordable, costly & very costly. The fee continues to be a constantly evolving actual fee, however, with the subsequent standards, it's far divided into over 4 lessons.

Thus the trouble of regressionproblem is transformed into classification dataset. The foremost weak point of choice bushes and Naive algorithm is the incapability to deal with numeric output lessons. The fee characteristic consequently needed to be divided into lessons comprising lots of price, however, this gives upward push to extra motives for "inaccuracies" [2].

The output facts of the classifier are broken up into education sets and take a look at set, 108 times and 28 test set times (general of 134 times).

2.2Dimensionality Reduction

It is an approach of range of random variables beneath neath consideration, via way of means of acquiring a set of key variables [3]. The better the range of capabilities, the extra hard it's far for the education set are to be visualised after which laboured on. Most of those capabilities regularly are linked, hence not needed. It is right here that classifiers of dimensionality are used[3].

Various varieties of classifiers for Reducing dimensionality i.e. choice of capabilities, extraction of capabilities.

2.3 Feature Selection

In choice of capabilities, we're interested in locating f of the k dimension that supply us the maximum detail and hence, discard the opposite dimensions (k − f) [4].

2.4 Feature Extraction

We are very readily interested in looking for a brand-new set of f dimension withinside of extraction of capabilities which might be versions of the unique d dimension, such as, PCA [4].

Algorithms for the choice of the apps are used right here. There are many approaches - Forrward selection and Backward selection.

2.5Forward Selections

In ahead choice, we begin without a variable and upload them one via way of means of one, including the only at every degree that maximum decreases the error, till any addition does now no longer lower the given error.

2.6 Backward Selections

In opposite choice we begin with all the variables and cast off them via a way of means of one, putting off at every step the only that maximum decreases the error (or best barely will increase it), earlier than any similar elimination drastically will increase the error [4].

3. Classification

Now let us undergo the very last degree that's class. As noted above, a separate take a look at the set which is being used for a particular classifier assessment and finding accuracies. Any class is correct if it can decide through measuring the range of sophistication samples effectively-recognized (genuine positives), the range of samples effectively recognized which aren't elegance members and the samples those have been both dislocated to the elegance (fake positives) or now no longer marked as elegance samples (fake negatives)[5]. Accuracy affords us with the statistics on how the per cent of times which might be categorized effectively. In statastics , it is given that the :

Accuracy = (Samples / Total Samples) X 100

4. Correlation

We use a function .corr() to find out the correlation of all the numerical between them. The screenshot depicts a heat map that shows the correlation of all the features.

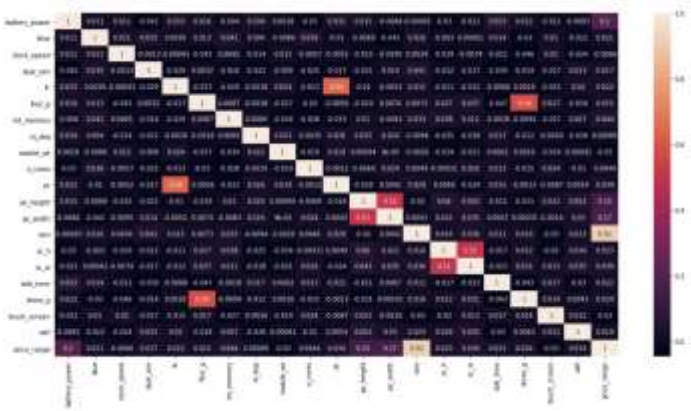

Fig-1 Heatmap of the correaltions

- Positive Correlation:- Positive correlation means if we have two features let's say A and B. If one A increases and B also increases or other way then we say they are positively related

- Negative Correlation:- Negative correlation means if we have two features let's say A and B. If one A increases and B decreases or other way then we say they are negatively related

- Zero Correlation:- If there is no relation between the two features we say they are not correlated.

5. Training Of Model Using SVM

The first set of rules to categorise and categorize the styles became provided via way of means became mistakes discount of categorized schooling records. Various of the strategies and many algorithms which have been imparted thus far to layout the classifications comply with this plan. In those strategies, designed category has a minute generalisation properties. If we remember the layout of classifying sample version as an optimisation hassle, lots of those strategies confront the hassle of neighbourhood optimization in an equation and are stuck withinside the entice of neighbourhood optimisation [6-8].

In 1965-"Vladimir" and "Pink" took a very crucial step inside the layout of the category [10-11]. He strongly set up statistical getting to know the idea and provided an aid vector system by it. The aid machines had the following properties-

1. Making of algorithim with the most extensions.
2. In taking the best factor out the whole function.
3. Automatically decide the choicest topology and shape for a classifier.
4. Modelling of nonlinear discriminate features the usage of the nonlinear cores and the of internal product.[9-11].

SVM are a set of rules that reveals a unique form of linear fashions and consequences in the most margin of the web page cloud. Maximising the given margins of web pages ought to bring about most separation among the class. The schooling factors to the most margin of the web cloud are known as aid vectors (factors). These are the handiest used to decide the reach among the classes.

If the records are linear and are separated, SVM train a linear system to supply the choicest degree that separate records without mistake with the most distance among the display screen and the schooling factors.

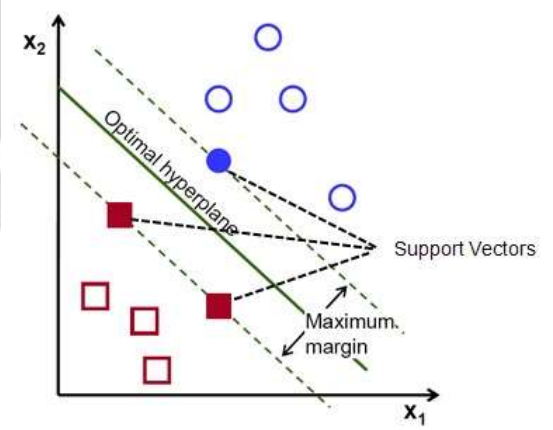

Fig-2 Working Of SVM Algorithim

When the given data is linear seperated,decision rules are explained by a optimised surfaces that separates binary decision classes are acording to given equation:

$$y = sign(\sum_{i=1}^{n} y_i a_i (X \, X) + b)$$

Fig-3 Formula of Linear Seperated Data

If data are linearly non-separable,

$$y = sign(\sum_{i=1}^{n} y_i a_i K(X \, X_i) + b)$$

Fig-4 Formula of Non-Linear Seperated Data

6. Standardization

A standard scalar is used to remove the mean and scale each feature to unit variance. When we train the model using SVM, we get an accuracy of 95% , which is good accuracy for a model to get trained.

```
from sklearn.preprocessing import StandardScaler
sc = StandardScaler()
X_train = sc.fit_transform(X_train)
X_test = sc.fit_transform(X_test)
```

```
from sklearn.svm import SVC
svm = SVC(kernel ='linear' , random_state= 0)
svm.fit(X_train,y_train)
```

```
SVC(kernel='linear', random_state=0)
```

```
from sklearn.metrics import accuracy_score
```

```
nos = svm.predict(X_test)
```

```
accuracy_score(y_test,nos)
```

```
0.95
```

Fig-5 Practical implementation of SVM Algorithim

7. Conclusion

This has finished achieving the most accuracy and decided on minimal, however maximum suitable capabilities. It is crucial to observe that during Forwarding choice via way of means of including beside the point or not needed capabilities to the facts. It decreases the performance of each classifier. When in backward choice, if we take away any crucial function from the facts set, the performance lows. The fundamental purpose of a low accuracy fee is a low quantity of times withinside the facts set. One extra factor that needs to additionally be taken into consideration whilst operating is that changing regression trouble into class trouble introduces extra error.

8. References

[1].Mobile data and specifications online available from ww.kaggle.com

[2].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

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

[4].S. Canbas, A. Cabuk, and S. B. Kilic. The Turkish case is the prediction of commercial bank failure via multivariate statistical analysis of financial structure. European Journal of Operational Research, 1:528–546, 2005.

[5].H. Frydman, E.I. Altman, and D. Kao. Introducing recursive partitioning for financial classification: The case of financial distress. Journal of Finance , 40(1):269–291, 1985.

[6].M. L. Marais, J. Patel, and M. Wolfson. The experimental design of classification models: An application of recursive partitioning and bootstrapping to commercial bank loan classifications. Journal of Accounting Research,22:87–114, 1984.

[7]. S. M. Bryant. A case-based reasoning approach to bankruptcy prediction modeling. Intelligent Systems in Accounting, Finance and Management, 6(3):195–214, 1997.

[8]. C. Park and I.Han. A case-based reasoning with the feature weights derived by analytic hierarchy process for bankruptcy prediction. Expert Systems with Applications, 23:255–264, 2002.

[9].K. S. Shin and Y. J. Lee. A genetic algorithm application in bankruptcy prediction modeling. Expert Systems with Applications, 23:312–328, 2002.

[10]. F. Varetto. A genetic algorithm application in bankruptcy prediction modeling. Journal of Banking and Finance, 22(10):1421–1439, 1998.

[11]. J. H. Min and Y. C. Lee. Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. Expert Systems with Applications, 28(4):603–614, 2005.