**Mobile Price Range Prediction**

**Sharad Tawade, Sagar Malik**

**Vinay Kumar**

**Data science trainee,**

**AlmaBetter**

**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.

Feature engineering is done to extract additional features from the data. Four statistical models are used to predict the price range

(a) Logistic regression

(b) Decision Tree

(c) Random Forest (RF).

(d) Gradient boosting machines

(e) Extreme Gradient Boosting (XGBoost)

(f) Ada-Boost

(g) Support Vector Classification

***Keywords:eda, machine learning, mobile price range, classification***

**1.Problem Statement**

### In the competitive mobile phone market companies want to understand sales data of mobile phones and factors which drive the prices. The objective is to find out some relation between features of a mobile phone(eg:- RAM, Internal Memory, etc) and its selling price. In this problem, we do not have to predict the actual price but a price range indicating how high the price is.

The main objective is to make a predictive model, which could help them in predicting

the mobile price demands proactively.

Attribute Information:

* **Battery\_power** - Total energy a battery can store in one time measured in mAh
* **Blue** - Has bluetooth or not
* **Clock\_speed** - speed at which microprocessor executes instructions
* **Dual\_sim** - Has dual sim support or not
* **Fc** - Front Camera mega pixels
* **Four\_g** - Has 4G or not
* **Int\_memory** - Internal Memory in Gigabytes
* **M\_dep** - Mobile Depth in cm
* **Mobile\_wt** - Weight of mobile phone
* **N\_cores** - Number of cores of processor
* **Pc** - Primary Camera mega pixels
* **Px\_height** - Pixel Resolution Height
* **Px\_width** - Pixel Resolution Width
* **Ram** - Random Access Memory in MegaBytes
* **Sc\_h** - Screen Height of mobile in cm
* **Sc\_w** - Screen Width of mobile in cm
* **Talk\_time** - longest time that a single battery charge will last when you are
* **Three\_g** - Has 3G or not
* **Touch\_screen** - Has touch screen or not
* **Wifi** - Has wifi or not
* **Price\_range** - This is the target variable with value of 0(low cost), 1(medium cost), 2(high cost) and 3(very high cost).

**2. Introduction**

Price is the most effective attribute of marketing and business. The very first question of a customer is about the price of items. All the customers are first worried and think “If he would be able to purchase something with given specifications or not”. So to estimate the 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 machines capable of answering the questions intelligently- nowadays is a very vast engineering field.

Machine learning provides us with the 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, cygwin, WEKA etc.

We here use many classifiers like Decision tree , Random Forest and many more. Different types of feature selection algorithms are available to select only the best features and minimize the dataset. This will reduce computational complexity of the problem. As this is an optimization problem so many optimization techniques are also used to reduce dimensionality of the dataset.

**3. Major Factors Affecting Mobile Price**

**i) Ram :** RAM is most correlated with mobile price, as RAM increases then mobile price also increases. RAM is very important as it holds the data on your phone. The more RAM your smartphone has, the more applications you can access quickly. Also, you can switch between multiple apps without facing any lag

**ii) Battery Power:** Battery is one of the most important parts of a mobile phone or a tablet as this is what keeps the device mobile. There are also limitations on size and weight as a big battery would make the device heavy and bulky.

**iii) Pixel Resolution:** Phone resolution principally depends on two significant factors, the screen size, and pixel count. In practice, this means that a smartphone with a higher resolution contains more pixels, and more pixels deliver the ability to display more visual information with greater clarity and detail.

**iv) Camera:** The study found that 92% of smartphone users use their phones to take pictures. 80% of users go on to send their photos elsewhere after they're snapped. Smartphone photography is more popular than Internet browsing, emailing, app downloading, and gaming.

**v) ROM/Storage:** Nowadays We required higher storage capacity in smartphones. We need to store photos, videos, apps, contacts, stream music and films, download emails, messages and webpages, and much more. It's no wonder all of these tasks take up so much storage.

**4. Steps Involved**

**I. Exploratory Data Analysis:**

Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, spot anomalies, test hypotheses and check assumptions with the help of summary statistics and graphical representations. It gives us a better idea of which feature behaves in which manner compared to the target variable.

**II. Data Cleaning:**

Our Data contains no null values and it is a multiclass classification problem, but our dataset is balanced with 4 classes.

**III. Features Selection:**

The feature selection process is based on a specific machine learning algorithm that we are trying to fit on a given dataset. It follows a greedy search approach by evaluating all the possible combinations of features against the evaluation criterion.

**IV. Feature Scaling:**

Feature scaling is essential for machine learning algorithms that calculate distances

between data. If not scale, the feature with a higher value range starts dominating when calculating distances.

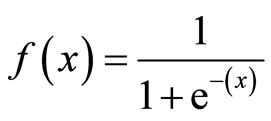
**V. Fitting Models:**

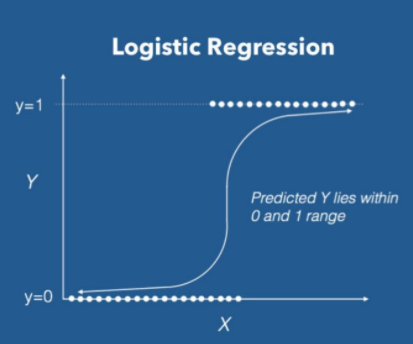
At first, we tried with basic logistic regression and then used a Decision tree Classifier, Random Forest Classifier, Gradient Boosting, XGB Model, Ada Boost and compared the results.

**5. Algorithms**

**I. Logistic Regression:**

Logistic regression is a classification algorithm used to assign observations to a discrete set of classes. Some of the examples of classification problems are Email spam or not spam, Online transactions Fraud or not Fraud, Tumor Malignant or Benign. Logistic regression transforms its output using the logistic sigmoid function to return a probability value. Sigmoid function is:





We can call a Logistic Regression a Linear Regression model but the Logistic Regression uses a more complex cost function, this cost function can be defined as the ‘Sigmoid function’ or also known as the ‘logistic function’ instead of a linear function.

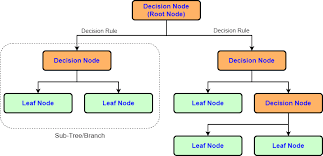
**II. Decision Tree:**

A decision Tree is a decision-making tool that uses a flowchart-like tree structure or is a model of decisions and all of their possible results, including outcomes, input costs and utility. Decision-tree algorithms fall under the category of supervised learning algorithms. It works for both continuous as well as categorical output variables. The branches/edges represent the result of the node and the nodes have either:

1. Conditions [Decision Nodes]

2. Result [End Nodes]

The branches/edges represent the truth/falsity of the statement and make a decision based on that in the example below which shows a decision tree.

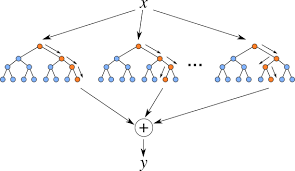


Decision tree classifiers are used successfully in many diverse areas. Their most important feature is the capability of capturing descriptive decision making knowledge from the supplied data. Decision trees can be generated from training sets.

**III. Random Forest:**

Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

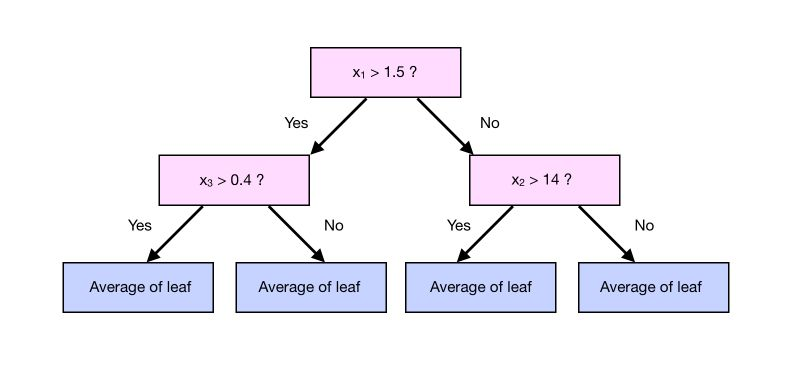
One of the most important features of the Random Forest Algorithm is that it can handle the data set containing continuous variables as in the case of regression and categorical variables as in the case of classification.



**III. XGBoost:**

To understand XGBoost we have to know gradient boosting beforehand.

* Gradient Boosting:



Gradient boosting is a type of machine learning boosting. It relies on the intuition that the best possible next model, when combined with previous models, minimizes the overall prediction error. The key idea is to set the target outcomes for this next model in order to minimize the error.

How are the targets calculated? The target outcome for each case in the data depends on how much changing that case's prediction impacts the overall prediction error.

**XGBoost** is one of the fastest implementations of gradient boosting trees. It does this by tackling one of the major inefficiencies of gradient boosted trees: considering the potential loss for all possible splits to create a new branch (especially if you consider the case where there are thousands of features, and therefore thousands of possible splits). XGBoost tackles this inefficiency by looking at the distribution of features across all data points in a leaf and using this information to reduce the search space of possible feature splits.

**6. Model Performance**

**I. Confusion Matrix:**

The confusion matrix is a table that summarizes how successful the classification models are at predicting examples belonging to various classes. One axis of the confusion matrix is the label that the model predicted, and the other axis is the actual label.

**II. Precision/Recall:**

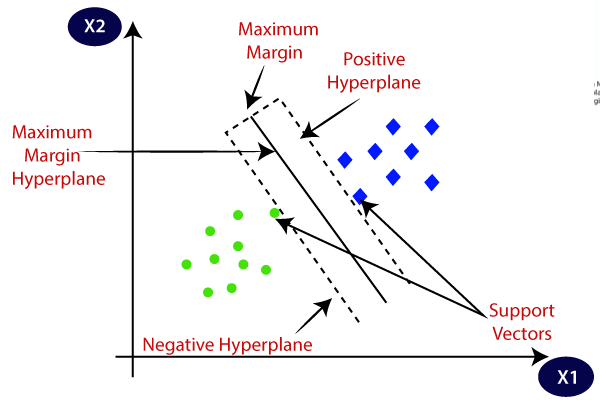
Precision is the ratio of correct positive predictions to the overall number of positive predictions: TP/TP+FP

Recall is the ratio of correct positive predictions to the overall number of positive examples in the set: TP/FN+TP.

**IV. Support Vector Machine:** Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems.

The goal of the 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. This best decision boundary is called a hyperplane.

SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called support vectors, and hence algorithm is termed a Support Vector Machine.

****

**III. Accuracy:**

Accuracy is given by the number of correctly classified examples divided by the total number of classified examples. In terms of the confusion matrix, it is given by: TP+TN/TP+TN+FP+FN

**IV. Area under ROC Curve(AUC):**

ROC curves use a combination of the true positive rate (the proportion of positive examples predicted correctly, defined exactly as recall) and the false positive rate (the proportion of negative examples predicted incorrectly) to build up a summary picture of the classification performance.

**7. HyperParameter Tuning**

Hyper-parameters are those sets of information that are used to control our parameters in order to get good results. We used Grid Search CV and Bayesian Optimization for hyperparameter tuning.

**Grid Search CV :**

It is the process of performing hyperparameter tuning in order

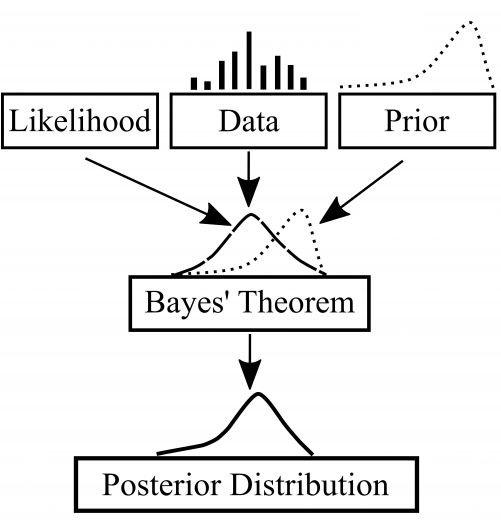
to determine the optimal values for a given model. As mentioned above, the performance of a model significantly depends on the value of hyperparameters. Note that there is no way to know in advance the best values for hyperparameters so ideally, we need to try all possible values to know the optimal values. Doing this manually could take a considerable amount of time and resources and thus we use GridSearchCV to automate the tuning of hyperparameters.

GridSearchCV is a function that comes in Scikit-learn’s(or SK-learn) model\_selection

package.So an important point here to note is that we need to have the Scikit-learn library installed on the computer. This function helps to loop through predefined hyperparameters and fit your estimator (model) on your training set. So, in the end, we can select the best parameters from the listed hyperparameters.

**Bayesian Optimization :**

Bayesian Hyperparameter optimization is a very efficient and interesting way to find good hyperparameters. In this approach, in naive interpretation way is to use a support model to find the best hyperparameters.A hyperparameter optimization process based on a probabilistic model, often Gaussian Process, will be used to find data from data observed in the later distribution of the performance of the given models or set of tested hyperparameters.



As it is a Bayesian process at each iteration, the distribution of the model’s performance in relation to the hyperparameters used is evaluated and a new probability distribution is generated. With this distribution, it is possible to make a more appropriate choice of the set of values that we will use so that our algorithm learns in the best possible way.

**8. Conclusion:**

The analysis was done on Mobile price range data with seven classification techniques namely: Logistic-Regression, Decision Tree, Random Forest,

XGBoost, KNN, Support Vector Classifiers and Neural Networks are used to predict the price range a mobile lies in. This statistical data analysis shows interesting outcomes in prediction methods and also in an exploratory analysis.

The experimental results prove that the Neural-Network model predicts the best range a mobile can lie in with the highest accuracy and AUC-ROC score and with less error rate compared to all the other algorithms used.