Mobile Price Classifictation

Medha B
MTECH (PG)
DSML
Pes University
Bangalore, Karnataka, India
medhab14126@gmail.com

Subham Budigas
MTECH (PG)
DSML
Pes University
Bangalore, Karnataka, India
subhamd1205@gmail.com

Debasish Rath
MTECH (PG)
DSML
Pes University
Bangalore, Karnataka, India
debasishrath301@gmail.com

***Abstract*—A mobile price classification has been a high-interest in the research area, and it requires a noticeable effort and knowledge of the field experts. To predict 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 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 best classifier 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.**

Keywords— Mobile Price Classification, Machine Learning, Classification, Naïve Bayes, Logistic Regression, kNN, AdaBoost, XGBoost, Gradient Boost, Random Forest, Decision Tree, Stack Generalization.

I. INTRODUCTION

Mobile phones have been one of the most essential components for us during the last decade. From Brick Phone to iPhone, technological advancement is seen in the sector of telecommunication. At present, mobile phones have different features according to the customer's choice and need. Among thousands of companies, offering their technology to the customers, choosing the appropriate phone for an individual has become an issue. Thus, the decision-making process becomes more challenging in recent times. In this research, a model is proposed which generates a price range of particular models, based on public requirements of different features or specification in a mobile phone. The price class i.e. below mid-range or above mid-range will let consumers compare between their need and satisfaction and thus choose the suitable mobile phone. There are many websites, YouTube channels and also individuals on Social Media who reviews mobile phones. Professional reviewers are appointed by different companies to act as a marketing policy of the company.

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 towards 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 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.

Mobile now a days is one of the most selling and purchasing device. Every day new mobiles with new version and more features are launched. Hundreds and thousands of mobile are sold and purchased on daily basis. So here the mobile price class prediction is a case study for the given type of problem i.e. finding optimal product.

II. LITERATURE SURVEY

Machine learning (ML) is the study of computer algorithms that improve automatically through experience and by the use of data. It is seen as a part of artificial intelligence. Machine learning algorithms build a model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to do so. Machine learning algorithms are used in a wide variety of applications, such as in medicine, email filtering, and computer vision, where it is difficult or unfeasible to develop conventional algorithms to perform the needed tasks.

A subset of machine learning is closely related to computational statistics, which focuses on making predictions using computers; but not all machine learning is statistical learning. The study of mathematical optimization delivers methods, theory and application domains to the field of machine learning.

Predicting the price of the mobile involves creating the models, which is trained on some training data and then can process additional data to make predictions. Various types of models have been used and researched for machine learning systems.

K-Nearest Neighbor is one of the simplest Machine Learning algorithms based on Supervised Learning technique-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories’-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.

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. 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.

Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems. It is mainly used in text classification that includes a high-dimensional training dataset.

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.

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. 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. As the name suggests, "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." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

Ada-boost or Adaptive Boosting is one of ensemble boosting classifier proposed by Yoav Freund and Robert Schapire in 1996. It combines multiple classifiers to increase the accuracy of classifiers. AdaBoost is an iterative ensemble method. AdaBoost classifier builds a strong classifier by combining multiple poorly performing classifiers so that you will get high accuracy strong classifier. The basic concept behind Adaboost is to set the weights of classifiers and training the data sample in each iteration such that it ensures the accurate predictions of unusual observations. Any machine learning algorithm can be used as base classifier if it accepts weights on the training set. Adaboost should meet two conditions:

XGBoost is well known to provide better solutions than other machine learning algorithms. In fact, since its inception, it has become the "state-of-the-art" machine learning algorithm to deal with structured data. XGBoost internally has parameters for cross-validation, regularization, user-defined objective functions, missing values, tree parameters, scikit-learn compatible API etc.

Stacked Generalization or stacking is an ensemble algorithm where a new model is trained to combine the predictions from two or more models already trained or your dataset.

The predictions from the existing models or sub models are combined using a new model, and as such stacking is often referred to as blending, as the predictions from sub-models are blended together.

III. WORKING PRINCIPLE

In this research, we use machine learning algorithms to find out price class of the mobiles. Furthermore, they also intended to find the best accuracy of the models used. We used different machine learning algorithms, such as Machine Learning, Classification, Naïve Bayes, Logistic Regression, kNN, AdaBoost, XGBoost, Gradient Boost, Random Forest, Decision Tree. A block diagram is given below to show the proposed model of mobile price prediction,

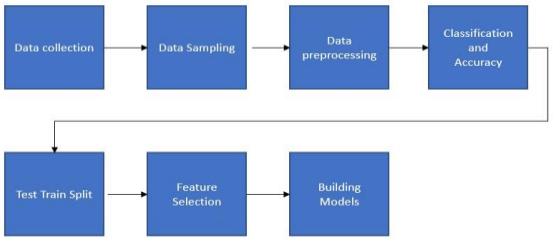

Fig. 1. Block diagram of the proposed model.

A. Data

Dataset is one of the most important factor in any kind of machine learning and predictions. In the research of mobile price prediction, the most technical and one of the toughest tasks is to find out the proper dataset applicable to the models. As the aim is to find out the proper dataset, we used a dataset that is collected from kaggle. There were two thousand rows of data in the dataset. The dataset had twenty one columns, which were 'battery power', 'clock_speed','dual_sim','fc','int_memory','m_dep','mobile_wt','n_cores','pc','px_height','px_width','ram','sc_h','sc_w','talk_time','three_g','touch_screen','wifi','price_range', all the columns are numerical data.

B. Data Sampling

Data sampling is a statistical analysis technique used to select, manipulate and analyze a representative subset of data points to identify patterns and trends in the larger data set being examined. It enables data scientists , predictive modelers and other data analysts to work with a small, manageable amount of data about a statistical population to build and run analytical models more quickly, while still producing accurate findings Sampling can be particularly useful with data sets that are too large to efficiently analyze in full -- for example, in big data analytics applications or surveys. Identifying and analyzing a representative sample is more efficient and cost-effective than surveying the entirety of the data or population.

An important consideration, though, is the size of the required data sample and the possibility of introducing a sampling error. In some cases, a small sample can reveal

the most important information about a data set. In others, using a larger sample can increase the likelihood of accurately representing the data as a whole, even though the increased size of the sample may impede ease of manipulation and interpretation.

C. Data Preprocessing

Data pre-processing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model. A real-world data generally contains noises, missing values, and maybe in an unusable format which cannot be directly used for machine learning models. Data pre-processing is required tasks for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model. There are several steps of data pre-processing, which are,

i. Getting the data
ii. Importing libraries
iii. Importing datasets
iv. Finding Missing Data
v. Encoding Categorical Data
vi. Splitting dataset into train and test split
vii. Feature scaling

To create a machine learning model, the first thing we required is a dataset as a machine learning model completely works on data. The collected data for a particular problem in a proper format is known as the dataset.

In order to perform data pre-processing using Python, we need to import some predefined Python libraries. These libraries are used to perform some specific jobs. There are three specific libraries that we will use for data pre-processing which are: Numpy, Matplotlib, Pandas.

We need to import the datasets which we have collected for our machine learning project and the next step of data pre-processing is to handle missing data in the datasets. If our dataset contains some missing data, then it may create a huge problem for our machine learning model. Hence it is necessary to handle missing values present in the dataset. Categorical data is data which has some categories such as, in our dataset that has to be encoded.

In machine learning data pre-processing, we divide our dataset into a training set and test set. This is one of the crucial steps of data pre-processing as by doing this, we can enhance the performance of our machine learning model.

Feature scaling is the final step of data pre-processing in machine learning. It is a technique to standardize the independent variables of the dataset in a specific range. In feature scaling, we put our variables in the same range and in the same scale so that no any variable dominate the other variable.

D. Classification and Accuracy

Classification We have used multiple algorithms for training the dataset. The amount of train data was increased gradually to see the accuracy of the algorithms. Classification is the task of choosing the correct class label for a given input. In the basic classification tasks, each input is considered in isolation from all other input, and the set of the labels is defined in advance. Mentioned earlier by using the built-in libraries to do all the classification process. To be specific, the following algorithms are used in the classification process, Naïve Bayes (NB), Logistic Regression, Decision Tree Classifier, KNN Classifier, Random Forest Machine learning algorithms are used to classify the sentences based on their polarity. However, the purpose of the research is to find out the best algorithm for sentiment analysis based on the accuracy of classification. The details of the algorithms are given below.

Naïve Bayes Classifier: A Naive Bayes classifier is an algorithm that uses Bayes' theorem to classify data. Naive Bayes classifiers assume strong, or naive, independence between attributes of data points. The classifier uses probability theory to classify. The main key insight of Bayes' theorem is that the probability of an event can be adjusted as new data is introduced. A Naïve Bayes classifier is called naïve because it assumes that all attributes of a data point under consideration are independent of each other. A Naïve Bayes model is easy to build and it has no complicated iterative parameters estimation which makes it particularly useful for very large datasets.

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (a form of binary regression). Mathematically, a binary logistic model has a dependent variable with two possible values, such as pass/fail which is represented by an indicator variable, where the two values are labeled "0" and "1". In the logistic model, the log-odds (the logarithm of the odds) for the value labeled "1" is a linear combination of one or more independent variables ("predictors"); the independent variables can each be a binary variable (two classes, coded by an indicator variable) or a continuous variable (any real value). The corresponding probability of the value labeled "1" can vary between 0 (certainly the value "0") and 1 (certainly the value "1"), hence the labeling; the function that converts log-odds to probability is the logistic function, hence the name. The unit of measurement for the log-odds scale is called a logit, from logistic unit, hence the alternative names. Analogous models with a different sigmoid function instead of the logistic function can also be used, such as the probability model; the defining characteristic of the logistic model is that increasing one of the independent variables multiplicatively scales the odds of the given outcome at a constant rate, with each independent variable having its own parameter; for a binary dependent variable this generalizes the odds ratio.

The k-nearest neighbours algorithm (k-NN) is a non-parametric classification method first developed by Evelyn Fix and Joseph Hodges in 1951, and later expanded by Thomas Cover. It is used for classification and regression. In both cases, the input consists of the k closest training examples in data set. The output depends on whether k-NN is used for classification or regression:
In k-NN classification, the output is a class membership. An object is classified by a plurality vote of its neighbours, with the object being assigned to the class most common among its k nearest neighbours (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbour.
In k-NN regression, the output is the property value for the object. This value is the average of the values of k nearest neighbours.

Decision Tree algorithm belongs to the family of supervised learning algorithms. Unlike other supervised learning algorithms, the decision tree algorithm can be used for solving regression and classification problems too. The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by learning simple decision rules inferred from prior data (training data). In Decision Trees, for predicting a class label for a record we start from the root of the tree. We compare the values of the root attribute with the record's attribute. On the basis of comparison, we follow the branch corresponding to that value and jump to the next node.

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean/average prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set. Random forests generally outperform decision trees, but their accuracy is lower than gradient boosted trees. However, data characteristics can affect their performance.
The first algorithm for random decision forests was created in 1995 by Tin Kam Ho using the random subspace method, which, in Ho's formulation, is a way to implement the "stochastic discrimination" approach to classification proposed by Eugene Kleinberg.

Stacking is a machine learning technique that takes several classification or regression models and uses their predictions as the input for the meta-classifier (final classifier) or meta-regressor (final regressor).

E. Train-Test Split

The train-test split procedure is used to estimate the performance of machine learning algorithms when they are used to make predictions on data not used to train the model.
It is a fast and easy procedure to perform, the results of which allow you to compare the performance of machine learning algorithms for your predictive modeling problem. Although simple to use and interpret, there are times when the procedure should not be used, such as when you have a small dataset and situations where additional configuration is required, such as when it is used for classification and the dataset is not balanced.

F. Feature Selection

Feature Selection is one of the core concepts in machine learning which hugely impacts the performance of your model. The data features that you use to train your machine learning models have a huge influence on the performance you can achieve. Irrelevant or partially relevant features can negatively impact model performance.
Feature selection and Data cleaning should be the first and most important step of your model designing.
In this post, you will discover feature selection techniques that you can use in Machine Learning. Feature Selection is the process where you automatically or manually select those features which contribute most to your prediction variable or output in which you are interested in.
Having irrelevant features in your data can decrease the accuracy of the models and make your model learn based on irrelevant features.
How to select features and what are Benefits of performing feature selection before modelling your data?
· Reduces Overfitting: Less redundant data means less opportunity to make decisions based on noise.
· Improves Accuracy: Less misleading data means modelling accuracy improves.
· Reduces Training Time: fewer data points reduce algorithm complexity and algorithms train faster.

G. Building Models

Model building is a hobby that involves the creation of physical models either from kits or from materials and components acquired by the builder. The kits contain several pieces that need to be assembled in order to make a final model. Most model-building categories have a range of common scales that make them manageable for the average person both to complete and display. A model is generally considered physical representations of an object and maintains accurate relationships between all of its aspects.

The model building kits can be classified according to skill levels that represent the degree of difficulty for the hobbyist. These include skill level 1 with snap-together pieces that do not require glue or paint; skill level 2, which requires glue and paint; and, skill level 3 kits that include smaller and more detailed parts. Advanced skill levels 4 and 5 kits ship with components that have extra-fine details. Particularly, level 5 requires expert-level skills.
Model building is not exclusively a hobbyist pursuit. The complexity of assembling representations of actual objects has become a career for several people. There are, for instance, those who build models to commemorate historic events, employed to construct models using past events as a basis to predict future events of high commercial interest.

	Model	Score Average
0	Gaussian Naive Bayes	0.927363
1	K-Nearest Neighbor	0.885754
2	Logistic Regression	0.992035
3	Decision Tree	0.851130
4	Random Forest	0.916743
5	XGB Classifier	0.970781
6	AdaBoost Classifier	0.969888
7	Gradient Boosting Classifier	0.966356
8	StackingClassifier	0.970781

Fig. 2. Average cross val scores of the models.

IV. Result and Discussion

In the proposed system, the Python programming language is used for implementation.

The following metrics can be used to evaluate the performance of classification models:

1. Confusion matrix
2. ROC

1. Confusion Matrix

Performance measure for classification problem, it is a table used to compare predicted and actual values of the target variable.

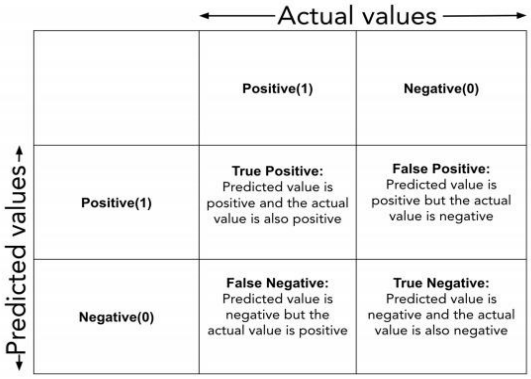

Fig. 3. Confusion Matrix.

Performance evaluation metrics:

i. Accuracy
ii. Precision
iii. Recall
iv. False Positive Rate
v. Specificity
vi. F1 score
vii. Kappa

i. Accuracy:

Accuracy is the fraction of predictions that out model got correct.

$$\text{Accuracy} = \frac{\text{number of correctly predicted records}}{\text{Total number of records}}$$

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

ii. Precision:

Precision is proportion of positive cases that were correctly predicted.

$$\text{Precision} = \frac{TP}{TP+FP}$$

iii. Recall:

Recall is the proportion of actual positive cases that were correctly predicted. Recall is also sometimes called True Positive Rate (TPR) or Sensitivity.

$$\text{Recall} = \frac{TP}{TP+FN}$$

iv. False Positive Rate:

False Positive Rate (FPR) is the proportion of actual negative cases that were predicted positive (incorrectly)

$$\text{FPR} = \frac{FP}{FP+TN}$$

$$\text{FPR} = 1 - \text{Specificity}$$

v. Specificity:

Specificity is the proportion of actual negative cases that were correctly predicted

$$\text{Specificity} = \frac{TN}{TN+FP}$$

vi. F1 Score:

F1 score is the harmonic mean of precision and recall values for a classification model. It is good measure if we want to find a balance between precision and recall or if there is uneven distribution of classes (either positive or negative class has way more actual instances than the other)

$$F_{1score} = \left(\frac{recall^{-1}+precison^{-1}}{2}\right)^{-1} = 2 . \frac{precision \; . \; recall}{precision + recall}$$

vii. Kappa:

Kappa statistic is a measure of interrater reliability or degree of agreement

$$p_o = \frac{\text{number of instances in agreement}}{\text{total instances}}$$

$$p_o = \frac{TP+TN}{TP+TN+FP+FN}$$

Identify the Important Features:

The bar plot is used to identify the important feature in the dataset.

The method `feature_importances_` returns the value corresponding to each feature which is defined as the ratio of total decrease in `Gini impurity` across every tree in the forest where the feature is used to the total count of trees in the forest. This is also called as, `Gini Importance`.

There is another `accuracy-based` method. It calculates the average decrease in the accuracy calculated on the out-of-bag samples, with and without shuffling the variable across all the trees in the random forest. The `out-of-bag` samples are the samples in the training dataset which are not considered while building a tree.

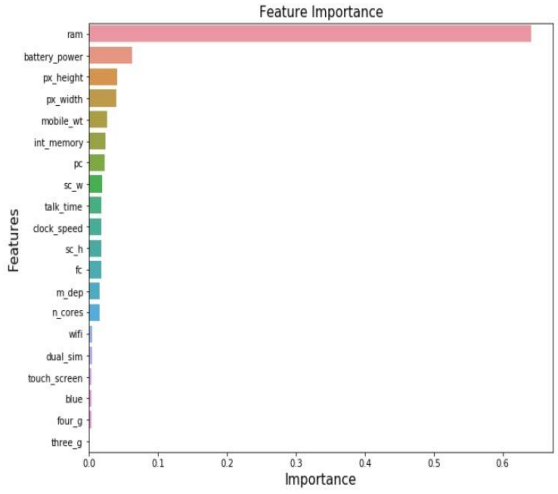

Fig. 4. Feature Importance

2. ROC

Receiver operating characteristics (ROC) curve is the TPR and FPR values change with different threshold values. ROC curve is the plot of TPR against the FPR values obtained at all possible threshold values.

Area under the ROC curve (AUC) is the measure of separability between the classes of target variables AUC increases as the separation between the classes increases Higher the AUC better the model.

The ROC curve for the best model is shown below

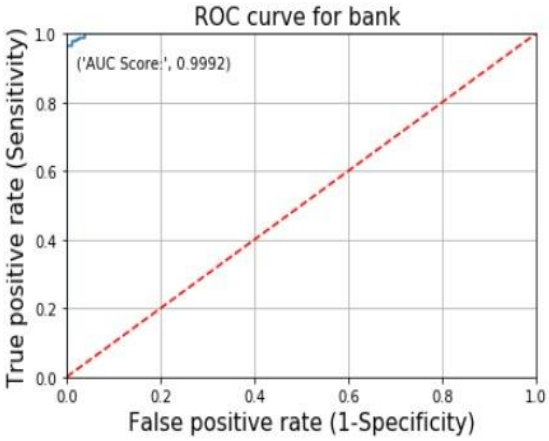

Fig. 5. Roc curve of the best model (Logistic Regression)

V. Conclusion

In today's world, uses of machine learning algorithms can play a significant role to understand people's opinion automatically and use that to improve certain product or service efficiently. This can be a field of interest for both researchers and entrepreneurs. In this research, analysing the price class of the mobile depending upon the features which satisfies the consumers need.

Naïve Bayes, Logistic Regression, Decision Tree , kNN, Random Forest, AdaBoost, Gradient Boost, XGBoost, Stacking Generalization algorithm are used in this model and accuracy of their performance are compared which showed that Logistic Regression and Stacking Generalization perform better with greater accuracy. Using multiple algorithms helped to understand which algorithm is more suitable for this system. Price range class of the product are given based on the average polarity obtained which can help the customer to know about certain products or services and take decisions accordingly. From this research work, it is found that, the cleaner the data, the better the accuracy of the result. In future, this model can be implemented as a web application so that while a customer is willing to buy a certain product, they can easily make a decision just by looking at the ratings. We want to improve the algorithms that are used in this research. In the future, the goal is to apply unsupervised machine learning in the model, such as the neural network. We are looking forward to making the decision-making process easier when it comes to selecting mobile phones for the users. We believe that this model will be user-friendly and accurate enough for users to use this on regular basis for judging mobile phones and save their time from looking to various websites to find the appropriate mobile that falls under their budget.

VI. REFERENCES

[1] https://ieeexplore.ieee.org/document/4421853

[2] https://ieeexplore.ieee.org/document/7041065

[3] https://ieeexplore.ieee.org/document/7041065

[4] https://ieeexplore.ieee.org/document/5916845

[5] https://ieeexplore.ieee.org/document/8710278

[6] Machine Learning with Python Cookbook: Practical Solutions from Preprocessing to Deep Learning Chris Albon (Author)

[7] https://www.kaggle.com/iabhishekofficial/mobile-price-classification

[8] Introduction to Machine Learning with Python: A Guide for Data Scientists 1st Edition by Andreas C. Müller.

[9] Deep Learning with Python 1st Edition by François Chollet.

[10] https://ieeexplore.ieee.org/document/6299258