Mobile Price Range Prediction

BY :- Ankur Nain

Introduction:

Mobile phones come in all sorts of prices, features, specifications and all. Price estimation and prediction is an important part of consumer strategy. Deciding on the correct price of a product is very important for the market success of a product. A new product that has to be launched, must have the correct price so that consumers find it appropriate to buy the product.

Exploratory data analysis:

exploratory data analysis (EDA) is an approach of analyzing data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling and thereby contrasts traditional hypothesis testing.

Exploratory data analysis tools:

Specific statistical functions and techniques you can perform with EDA tools include:
Clustering and dimension reduction techniques, which help create graphical displays of high-dimensional data containing many variables. Univariate visualization of each field in the raw dataset,

with summary statistics. Bivariate visualizations and summary statistics that allow you to assess the relationship between each variable in the dataset and the target variable you're looking at. Multivariate visualizations, for mapping and understanding interactions between different fields in the data. K-means Clustering is a clustering method in unsupervised learning where data points are assigned into K groups, i.e. the number of clusters, based on the distance from each group's centroid. The data points closest to a particular centroid will be clustered under the same category. K-means Clustering is commonly used in market segmentation, pattern recognition, and image compression. Predictive models, such as linear regression, use statistics and data to predict outcomes.

Types of exploratory data analysis:

There are four primary types of EDA:

Univariate non-graphical :

This is simplest form of data analysis, where the data being analyzed consists of just one variable. Since it's a single variable, it doesn't deal with causes or relationships. The main purpose of univariate analysis is to describe the data and find patterns that exist within it.

Univariate graphical :

Non-graphical methods don't provide a full picture of the data. Graphical methods are therefore required. Common types of univariate graphics include: Stem-and-leaf plots, which show all

data values and the shape of the distribution. Histograms, a bar plot in which each bar represents the frequency (count) or proportion (count/total count) of cases for a range of values. Box plots, which graphically depict the five-number summary of minimum, first quartile, median, third quartile, and maximum.

Multivariate nongraphical :

Multivariate data arises from more than one variable. Multivariate non-graphical EDA techniques generally show the relationship between two or more variables of the data through cross-tabulation or statistics.

Multivariate graphical :

Multivariate data uses graphics to display relationships between two or more sets of data. The most used graphic is a grouped bar plot or bar chart with each group representing one level of one of the variables and each bar within a group representing the levels of the other variable. Other common types of multivariate graphics include: Scatter plot, which is used to plot data points on a horizontal and a vertical axis to show how much one variable is affected by another. Multivariate chart, which is a graphical representation of the relationships between factors and a response. Run chart, which is a line graph of data plotted over time. Bubble chart, which is a data visualization that displays multiple circles (bubbles) in a two-dimensional plot. Heat map, which is a graphical representation of data where values are depicted by color.

Problem statement

•In the competitive mobile phone market companies want to understand sales data of mobile phones and Factors which drives the prices.

•The objective is to find out some relation between features of a mobile phone and its selling price in this problem we do not have to predict theatrical process but a price range indicating how high the price is.

Problem:

The data contains information regarding mobile phone features, specifications etc and their price range. The various features and information can be used to predict the price range of a mobile phone.

The data features are as follows:

- Battery Power in mAh
- Has BlueTooth or not
- Microprocessor clock speed
- The phone has dual sim support or not
- Front Camera Megapixels
- Has 4G support or not
- Internal Memory in GigaBytes
- Mobile Depth in Cm
- Weight of Mobile Phone
- Number of cores in the processor
- Primary Camera Megapixels
- Pixel Resolution height
- Pixel resolution width
- RAM in MB
- Mobile screen height in cm
- Mobile screen width in cm
- Longest time after a single charge
- 3g or not
- Has touch screen or not
- Has wifi or not

Output Files:

model.h5: - Model contains information about the predictions of the train set, such as 0(low),high(1),very high(2).
confusion_matrix.txt : - Contains information about the classified emotions of the test set.

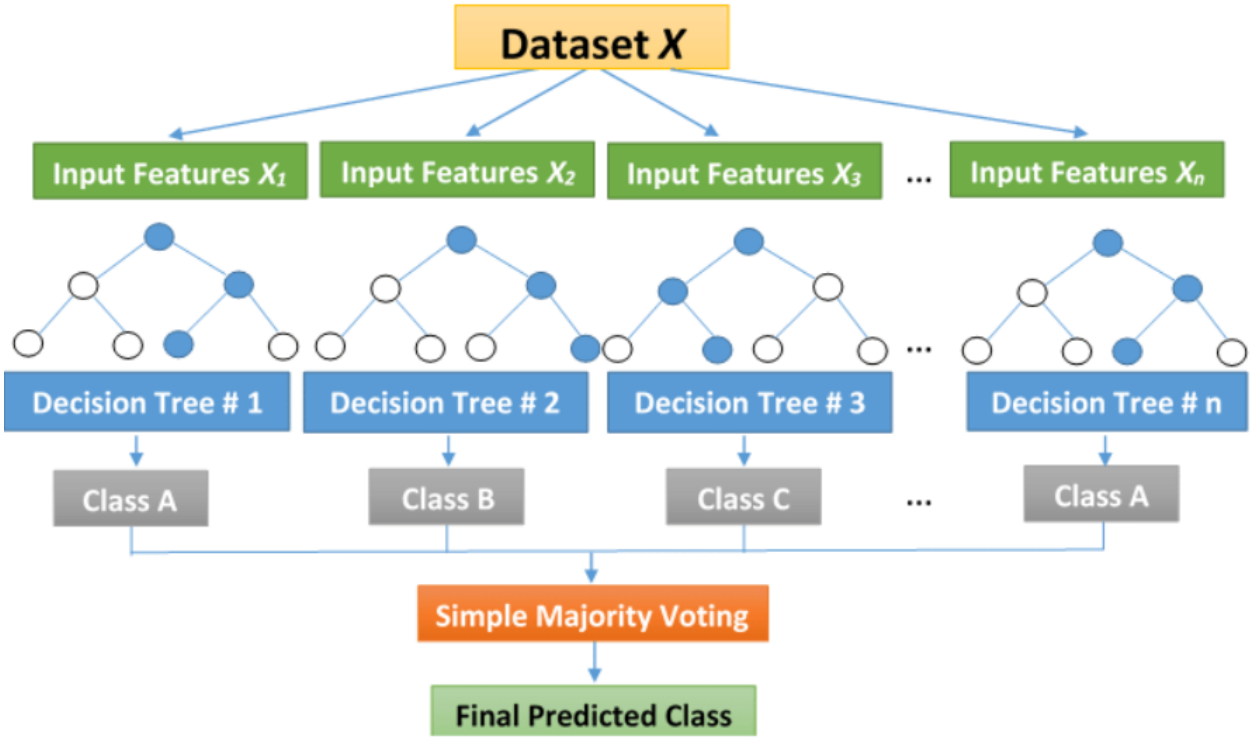
Source Directories:

Dataset - Includes all dataset for the training phase and testing phase of the model in the csv format.

📖 Random Forest Classification:

Random forest classifier creates a set of decision trees from randomly selected subset of training set.

It then aggregates the votes from different decision trees to decide the final class of the test object



Random forest classifier creates a set of decision trees from randomly selected subset of training set. It then aggregates the votes from different decision trees to decide the final class of the test object.Ensembled algorithms are those which combines more than one algorithms of same or different kind for classifying objects. For example, running prediction over Naive Bayes, SVM and Decision Tree and then taking vote for final consideration of class for test object.Basic parameters to Random Forest Classifier can be total number of trees to be generated and decision tree related parameters like minimum split, split criteria etc.

📖 XGboost

XGBoost stands for “Extreme Gradient Boosting”. XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It implements Machine Learning

algorithms under the Gradient Boosting framework. It provides a parallel tree boosting to solve many data science problems in a fast and accurate way.

DATA DESCRIPTION:

0	Battery_power	int64
1	blue	int64
2	clock_speed	float64
3	dual_sim	int64
4	fc	int64
5	four_g	int64
6	int_memory	int64
7	m_dep	float64
8	mobile_wt	int64
9	n_cores	int64
10	pc	int64
11	px_height	int64
12	px_width	int64
13	ram	int64

14	sc_h	int64
15	sc_w	int64
16	talk_time	int64
17	three_g	int64
18	touch_screen	int64
19	wifi	int64
20	price_range	int64

Dtypes:-

1.float64(2)

2. int64(19)

Features Breakdown:

The data contains information regarding mobile phone features,specification etc and their price range the various features and information can be used to predict the price range of a mobile phone .

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 - intern 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 – Pixels resolution height.
Px_weight – pixels resolution width.
Ram – random Access Memory in Mega Bytes.
Sc_h – Screen height of mobile in cm.
Talk_time – longest time that a single battery charge will 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 varible with values of 0(low cost),1(medium cost),2(high cost) and 3 (very high cost).

Conclusion:

•from EDA we can see that here are mobile phones in 4 price ranges. the number of element is almost similar.
•half the device have Bluetooth and half don't.
•costly phones are lighter.
•RAM battery power ,pixels played more sighnification role in deciding the price range of mobile phone.
•form all the above experiment we can conclude that logistic regression and XGbosting with using hyperparameters we got the best results.

•The accuracy and performance of the model is evaluated by using confusion matrix