Mobile Price Range Prediction

# By Harshad Savle

**Data Science Trainee AlmaBetter, Bangalore**

# Abstract:

To predict “if the mobile with given feature will be Economical or Expensive” is the main motive of this project. Different feature selection algorithms are used to identify and remove less important and redundant features and have minimum computational complexity.

Different feature selection algorithms are used to identify and remove less important and redundant features and have minimum computational complexity. 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.

***Keywords: EDA, Correlation, Decision Tree , Random Forest, XGboost, Classification, Forecasting***

# Problem Statement:

In the competitive mobile phone market companies want to understand sales data of mobile phones and factors which drive the price. The objective is to find out some relation between features of a mobile phone (ex:- 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.

# Introduction:

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.

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.

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. The same work can be done to estimate real price of all products like cars, bikes , generators, motors, food items, medicine etc.

Many features are very important to be considered to estimate price of mobile. For example RAM of the mobile. Battery capacity is also very important in todays busy schedule of human being. Size and thickness of the mobile are also important decision factors. Internal memory, Camera pixels, and video quality must be under consideration. And so is the list of many features based upon those, mobile price is decided. So we will use many of above mentioned features to classify whether the mobile would be very economical, economical, expensive or very expensive.

# Approach:

The approach followed here is to first check the sanctity of the data and then understand the features involved. The events followed were in our approach:

## Understanding the problem statement and the datasets

* **Data cleaning and preprocessing** finding null values and imputing them with appropriate values. Converting categorical values into appropriate data types and merging the datasets provided to get a final dataset to work upon.
* **Exploratory data analysis-** of categorical and continuous variables against our target variable.
* **Data manipulation-** feature selection and engineering, feature scaling, outlier detection and treatment and encoding categorical features.
* **Modeling**- The baseline model- Logistic regression was chosen considering our features were mostly categorical with few having continuous importance.
* **Model Performance and Evaluation**

## Price range prediction

* **Conclusion and Recommendations**

# Understanding the Data:

First step involved is understanding the data and getting answers to some basic questions like; What is the data about? How many rows or observations are there in it? How many features are there in it? What are the data types? Are there any missing values? And anything that could be relevant and useful to our investigation. Let’s understand the dataset first and the terms involved before proceeding further.

Our dataset consists of two csv files, the first consists of historical data with 2000 rows or observations and 21 columns with no null values.

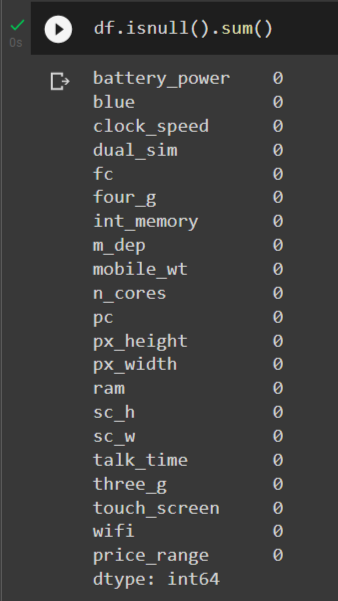
The data types were of integer, float and object in nature.

Let’s define the features involved:

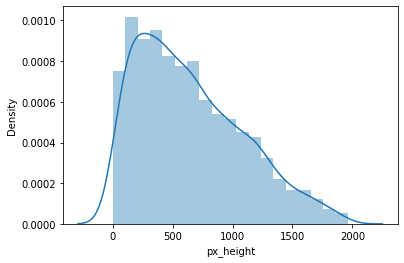
* **Battery\_power** - Total energy a battery can store in one time measured in mah
* **Blue** – Has Bluetooth or not
* **Clock\_speed** - speed which microprocessor executes instructions
* **Dual\_sim** - has dual sim support or not
* **Fc** – Front camera mega pixels.
* **Battery\_power** -Total energy 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 Data description
* **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 Mega Bytes
* **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.

**Data Cleaning and Preprocessing:** Handling missing values is an important skill in the data analysis process. If there are very few missing values compared to the size of the dataset, we may choose to drop rows that have missing values. Otherwise, it is better to replace them with appropriate values. It is necessary to check and handle these values before feeding it to the models, so as to obtain good insights on what the data is trying to say and make great characterisations and predictions which will in turn help improve the business's growth.

* The historical records of mobile dataset had no null values.

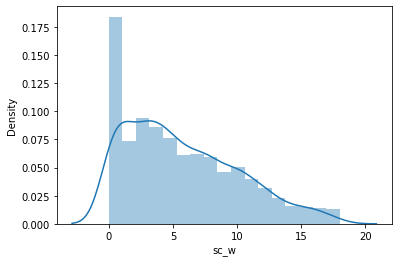


* Px\_height -Px\_height is the pixel resolution height of the camera in the given dataset. Px\_height distribution plot shows pixel resolution height generally present in the mobile camera.



* Right skewed distributions occur when the long tail is on the right side of the distribution also called as positive skewed distribution which essentially suggests that there are positive outliers far along which influences the mean. It seems like most of the values of the Skewed towards right.
* ‘Sc\_w ’ - sc\_w is the screen width measured in ‘cm’ given in the dataset. Sc\_w distribution plot shows screen width generally the phones have.

Right skewed distributions occur when the long tail is on the right side of the distribution also called as positive skewed distribution which essentially suggests that there are positive outliers far along which influences the mean. It seems like most of the values of the Skewed towards right.



* It seems like most of the values of the Sc\_w are towards the left and the distribution is skewed on the right. mean is more robust to outlier effect hence mean was imputed in the zero values.

# Exploratory Data Analysis:

Exploratory data analysis is a crucial part of data analysis. It involves exploring and analyzing the dataset given to find out patterns, trends and conclusions to make better decisions related to the data, often using statistical graphics and other data visualization tools to summarize the results. The visualization tools involved in the investigation are python libraries- matplotlib and seaborn.

The goal here is to explore the relationships of different variables with ‘price range’ to see what factors might be contributing to the high and low price range.

## Approach:

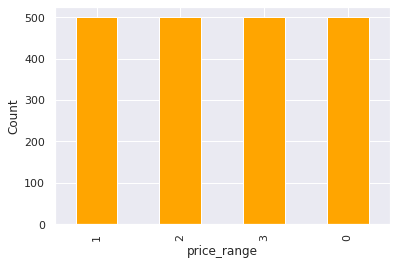
There are two kinds of features in the dataset: Categorical and Non Categorical Variables. Categorical- A categorical variable is a variable that can take on one of a limited, and usually fixed, number of possible values putting a a particular category to the observation. Non Categorical- A non-categorical or continuous variable is a variable whose value is obtained by measuring, i.e., one which can take on an uncountable set of values. Both of them are analyzed separately. Categorical data is usually analyzed through count plots and barplots in accordance with the target variable and that is what is done here too. On the other hand Numeric or Continuous variables were analyzed through distribution plots, box plots and scatterplots to get useful insights.

## Hypotheses

Just by observing the head of the dataset and understanding the features involved in it, the following hypotheses could be framed:

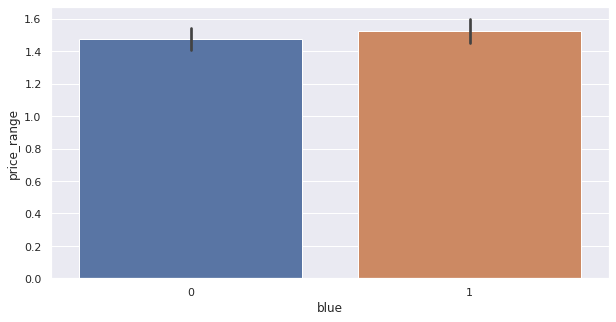
* + - There are mobile phones in 4 price ranges. The number of elements is almost similar.
    - The feature called “ram” will be most positively correlated variable with the dependent feature called “price range”
    - The front camera and the primary camera pixel size should be having a positive correlation with price range.
    - Half of the devices have blue tooth and half of the devices don’t hBluetooth.

## Price range distribution:



Here it can be seen that the price range is uniformly distributed all over the range. This validates the hypothesis about this feature.

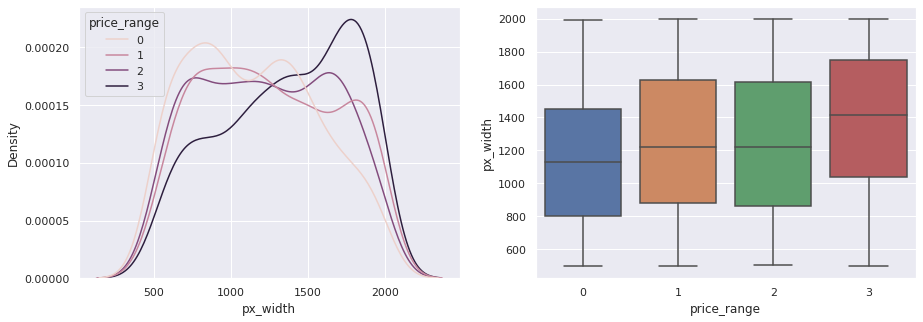
## Bluetooth Distribution :



In the above chart half devices have Bluetooth and half the devices don’t have Bluetooth.**.**

**Pixel Width:**

There is not continuous increase in pixel width as we move from low cost to high cost mobiles with ‘Medium cost’ and ‘High cost’ has almost equal pixel width. May be pixel size can say that it would be a driving factor in deciding price range.

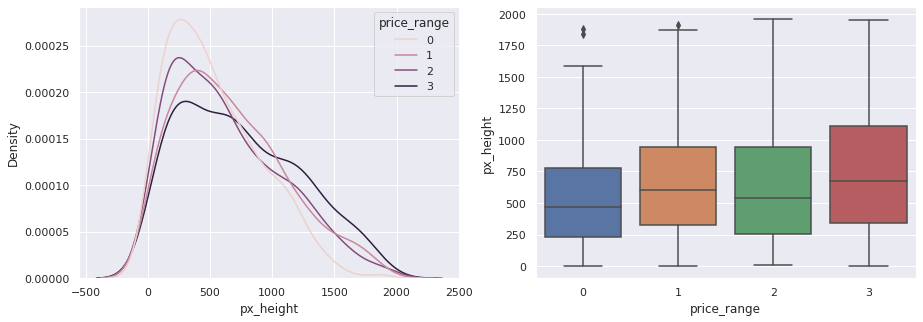
****

## Ram Distribution :

## As we can see from the below scatter plot as ram is increasing the price range is also increasing.

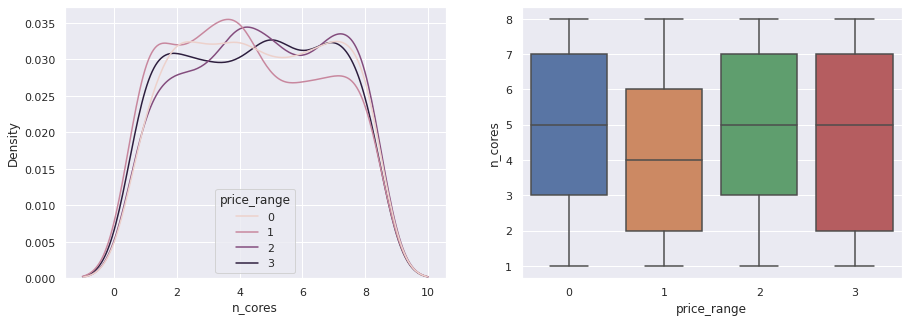
## 

Pixel height :



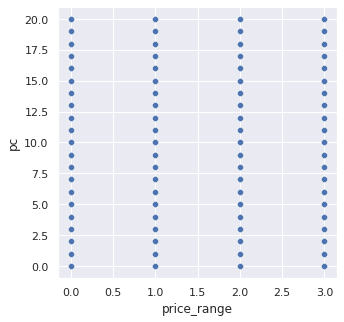
The pixel height is almost similar as we move from low cost to very high cost. Little variation pixel height.

**N cores (number of core in processor) :**

****

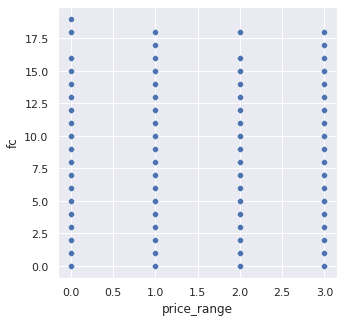
Here in primary camera megapixel are showing a little variation along with the target categories, which is good sign for prediction.

## Primary camera :

****

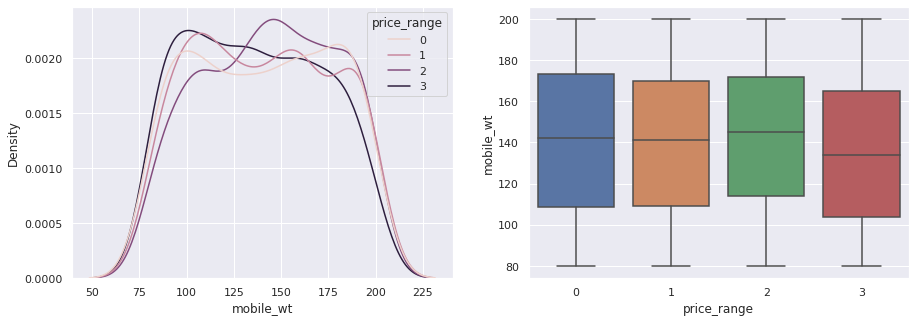
Primary camera is not so useful for us to make prediction is almost similar along all the price ranges.

## Front camera :



Both front and Primary camera are not so useful for us to make prediction is almost similar along all the price ranges.

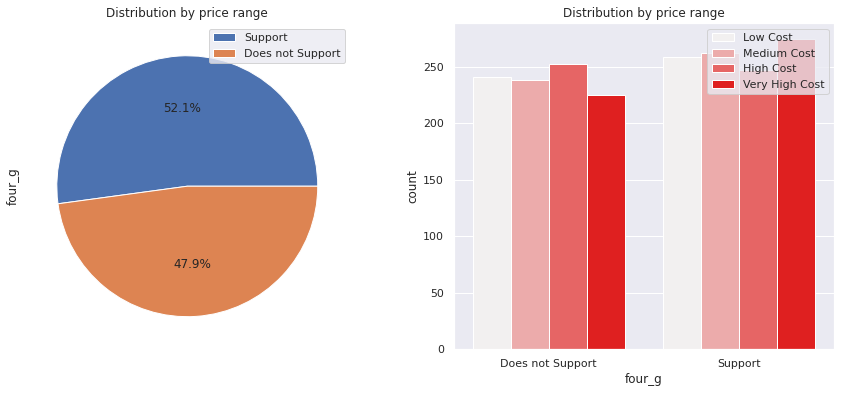
## Mobile weight distribution :



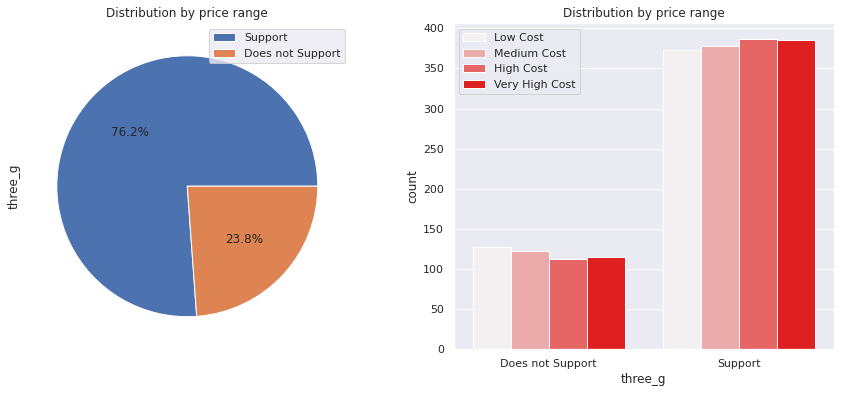
The above distribution shows mobile

weight distribution and with four categories and we can see that the costly phones are lighter than cheaper one.

## Binary features against price range :



From above distribution we can see that almost half mobile phones support 4 g and half do not support 4g connectivity.

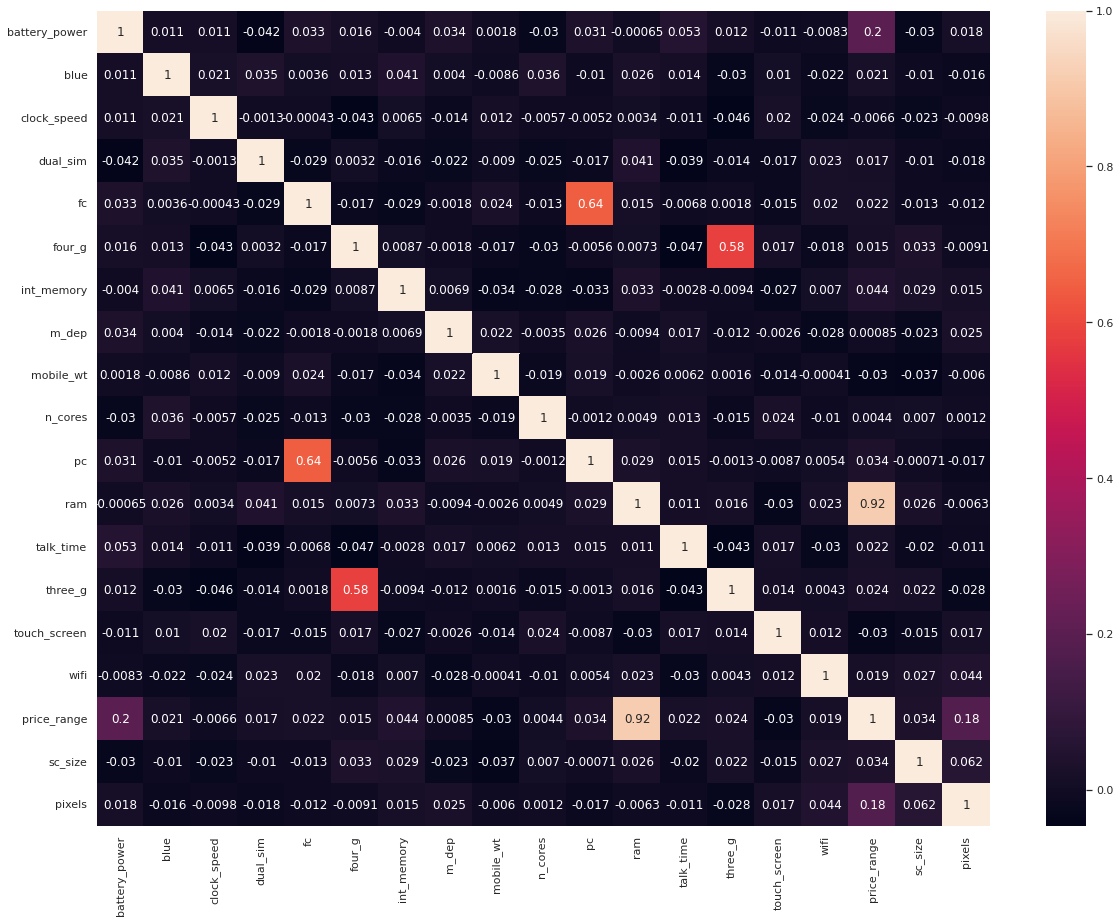


Above we can see that feature three g play an important role in prediction of price ranges.

## Correlation:

Correlation is a statistical term used to measure the degree in which two variables move in relation to each other. A perfect positive correlation means that the correlation coefficient is exactly 1. This implies that as one variable moves, either up or down, the other moves in the same direction. A perfect negative correlation means that two variables move in opposite directions, while a zero correlation implies no linear relationship at all.

By checking the correlation the factors affecting sales can be figured out.



We didn't find any strong correlation between independant variables but we found some correlation with our dependent feature which is a good sign for our model.

* Ram and price range shows high correlation which is a good sign, it signifies that ram will play major deciding factor in estimating the price range.
* As Pixel height was increases, pixel width was also increases that means the overall pixels in the screen. So we replaced two features with one feature. Front camera megapixels and primary camera megapixel are different entities despite of showing collinearity. so we kept them as they are.
* Also,There is some collinearity in feature pairs (pc,fc) and (px\_width,px\_height). Both correlations are justified since there are good chances that if primary camera of a phone is good quality, then front camera would also be good quality.

# Data Manipulation:

Data manipulation involves manipulating and changing our dataset before feeding it to various regression machine learning models. This involves keeping important features, outlier treatment, feature scaling and creating dummy variables if necessary.

## Feature Engineering:

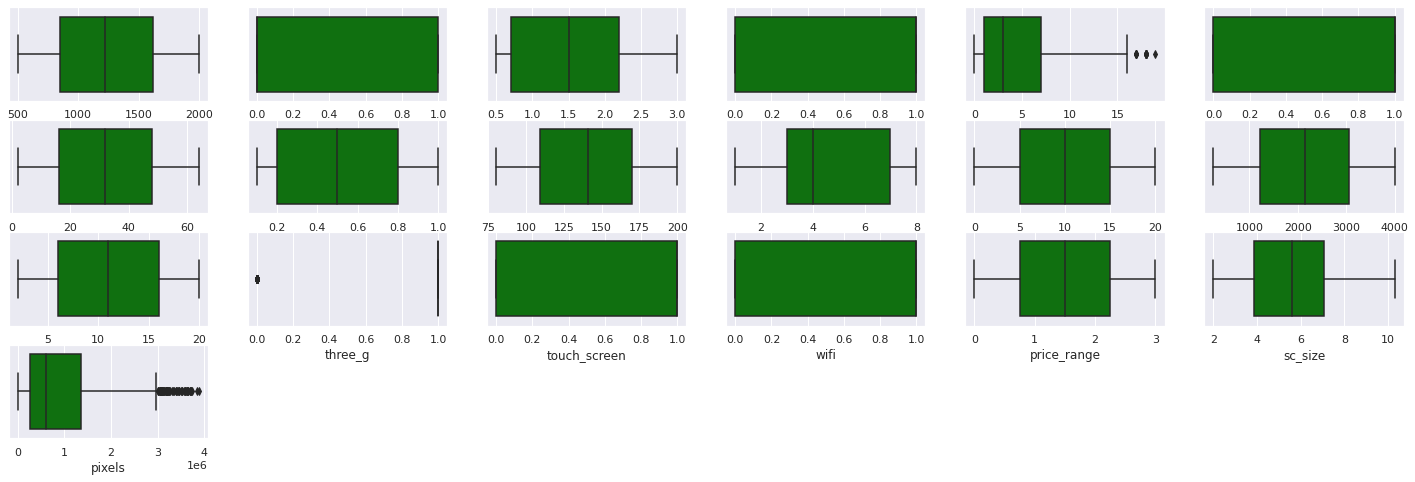
* There was feature called sc\_ h and sc\_w , sc\_h indicates screen height and sc\_w indicates screen width hence we made combination of feature named with sc\_size i.e. screen size.
* We already made useful feature with combination of screen height and screen width hence there was no sense in keeping those feature and they will create problem in prediction so we dropped those features.

## Outlier Detection:

In statistics, an outlier is a data point that differs significantly from other observations. Outliers can occur by chance in any distribution, but they often indicate either measurement error or that the population has a heavy-tailed distribution.

Z-score is a statistical measure that tells you how far a data point is from the rest of the dataset. In a more technical term, Z-score tells how many standard deviations away a given observation is from the mean.

z = (x-mean)/standard deviation



## From above feature we can see that there is no outlier present in the dataset.

## Feature Scaling:

Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is done to prevent biased nature of machine learning algorithms towards features with greater values and scale. The two techniques are: Normalization: is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min- Max scaling. [0,1]



Standardization: is another scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation. [-1,1]



Normalization of the continuous variables was done further.

**Modeling:**

Factors affecting in choosing the model:

Determining which algorithm to use depends on many factors like the problem statement and the kind of output you want, type and size of the data, the available computational time, number of features, and observations in the data, to name a few.

The dataset used in this analysis has:

* + A multivariate time series relation with sales and hence a linear relationship cannot be assumed in this analysis. This kind of dataset has patterns such as camera pixels, etc which would most likely be considered as outliers in simple linear regression.

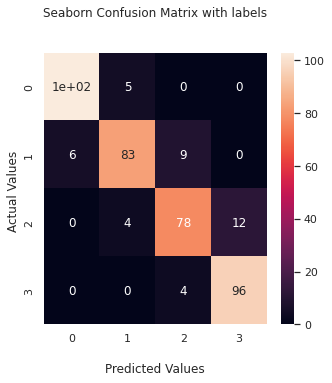
Having X columns with 30% continuous and 70% categorical features.

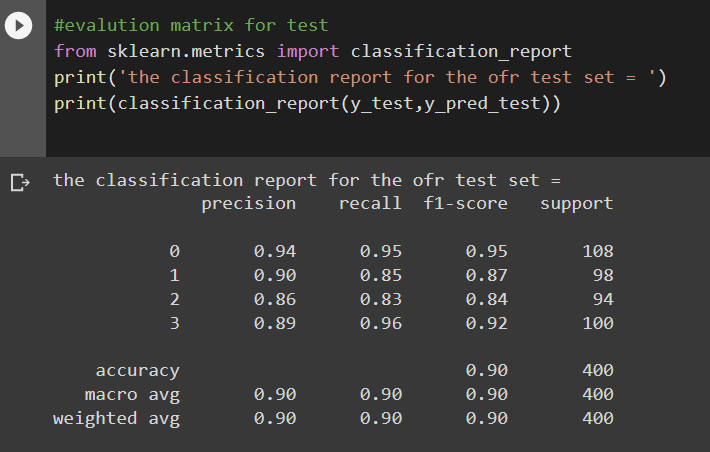
## Train-Test Split:

In machine learning, train/test split splits the data randomly, as there’s no dependence from one observation to the other. That’s not the case with time series data. Here, it’s important to use values at the rear of the dataset for testing and everything else for training.

**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, Tumour Malignant or Benign. 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. The hypothesis of logistic regression tends it to limit the cost function between 0 and 1. Therefore linear functions fail to represent it as it can have a value greater than 1 or less than 0 which is not possible as per the hypothesis of logistic regression. Making predictions with a logistic regression model is as simple as plugging in numbers into the logistic regression equation and calculating a result.

The confusion matrix shown below is for test dataset of in logistic regression model and the precision and recall results.

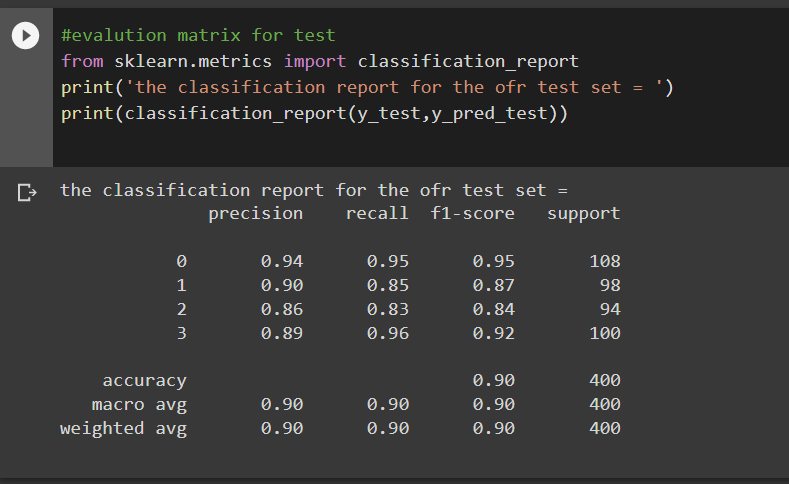




## Random Forest:

Random forests are an ensemble learning method for classification and regression that operates by constructing a multitude of decision trees at training time. For regression tasks, the output of the random forest is the average of the results given by most trees.

In simple terms, random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction.



**Random Forest Hyperparameters:**

* max\_depth- The max\_depth of a tree in Random Forest is defined as the longest path between the root node and the leaf node
* min\_sample\_split- a parameter that tells the decision tree in a random forest the minimum required number of observations in any given node in order to split it.
* The default value of the minimum\_sample\_split is assigned to 2. This means that if any terminal node has more than two observations and is not a pure node, we can split it further into subnodes.
* max\_leaf\_nodes- This hyperparameter sets a condition on the splitting of the nodes in the

tree and hence restricts the growth of the tree. If after splitting we have more terminal nodes than the specified number of terminal nodes, it will stop the splitting and the tree will not grow further.

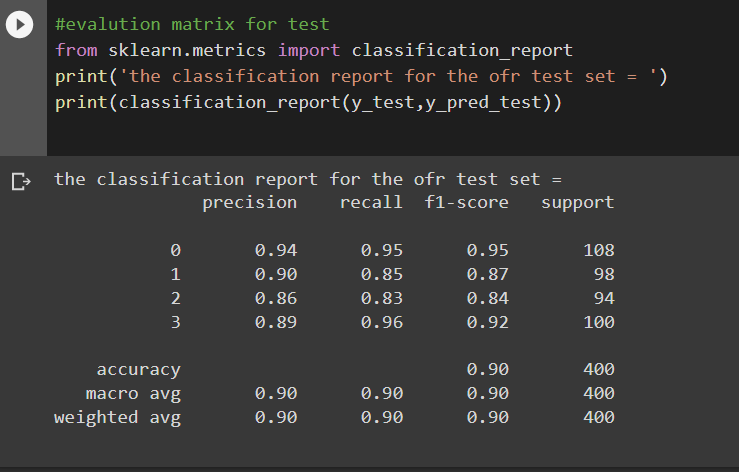
* + min\_samples\_leaf- This Random Forest hyperparameter specifies the minimum number of samples that should be present in the leaf node after splitting a node.
  + n\_estimators- the number of trees
  + max\_sample (bootstrap sample)-The max\_samples hyperparameter determines what fraction of the original dataset is given to any individual tree.
  + max\_features- This resembles the number of maximum features provided to each tree in a random forest.

Randomized searchcv searches on hyper parameters to fit and score various models and get the best estimator. In contrast to GridSearchCV, not all parameter values are tried out, but rather a fixed number of parameter settings is sampled from the specified distributions. The number of parameter settings that are tried is given by n\_iter.

## Random Forest Hyperparameter Tuned Model:

The maximum accuracy was seen in the tuned Random Forest model with the accuracy of 90% which was only 0.001% improved from a simple random forest model.

This indicates that all the trends and patterns that could be captured by these models without overfitting were done and the maximum level of performance achievable by the model was achieved.

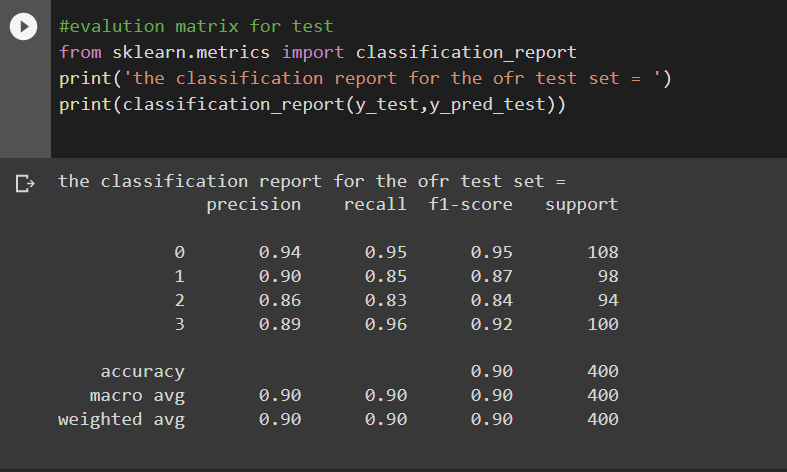


# Decision Tree:

In computer science, Decision tree learning uses a decision tree (as a predictive model) to go from observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves). It is one of the predictive modelling approaches used in statistics, data mining and machine learning. Tree models where the target variable can take a discrete set of values are called classification trees; in these tree structures, leaves represent class labels and branches represent conjunctions of features that

lead to those class labels. Decision trees where the target variable can take continuous values (typically real numbers) are called regression trees. In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. In data mining, a decision tree describes data (but the resulting classification

tree can be an input for decision making). Decision Tree Classifier, repetitively divides the working area(plot) into sub part by identifying lines. Decision tree at every stage selects the one that gives best information gain.



## Decisiontree with hyperparameter:

## Below results shows the classification results with grid search hyperparameter tuning technique.

## 

## XG Boost:

## XGBoost is and implementation of gradient boosted decision trees. In this algorithm, decision trees are created in sequential form. Weights play an important role in XGboost. Weights are assigned to all the independent variables which are then

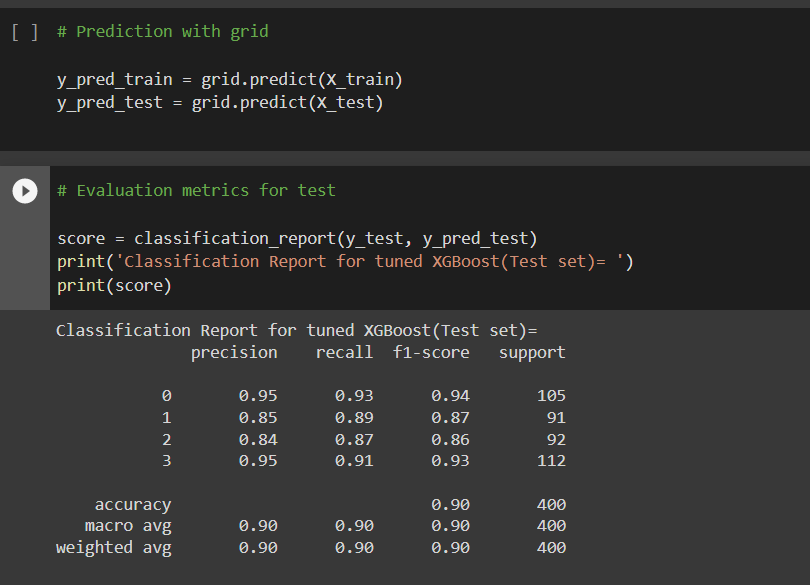
## fed into the decision tree which predicts results. The weight of variables predicted wrong by the tree is increased and these variables are then fed to the second decision tree. These individual classifiers then ensemble to give a strong and more precise model. It can work on regression, classification, ranking, and user-defined prediction problems.

## 

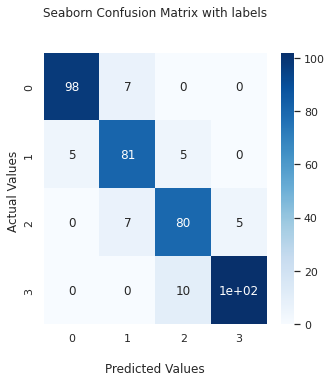
## The above result shows precision, recall and f1- score basemodel of XG boost algorithm.

## XGBoost Hyperparameter Tuned Model:

* n\_estimators- The number of trees
* learning rate- step size shrinkage used in update to prevents overfitting. After each boosting step, we can directly get the weights of new features, and eta shrinks the features weights to make the boosting process more conservative.
* gamma- Minimum loss reduction required to make a further partition on a leaf node of the tree. The larger gamma is, the more conservative the algorithm will be.
* subsample- Subsample ratio of the training instances. Setting it to 0.5 means that XGBoost would randomly sample half of the training data prior to growing trees. and this will prevent overfitting. Subsampling will occur once in every boosting iteration.



We can see the results of XGBoost with hyper tuned model.



**Conclusion**:

* + From the exploratory data analysis we can easily see that here are mobile phones in 4 different price ranges.
  + The number of the elements is almost similar.
  + Half the devices have Bluetooth, and half don’t.
  + Costly phones are lighter and ram, battery power, pixels played more significant role in deciding the price range of mobile phone.
  + There is a gradual increase in battery as the price range while moving from low cost to very high cost.
  + Form all the above experiments we can conclude that logistic regression and, XGBoosting with hyperparameters we got the best results.

## References:

* + GeeksforGeeks
  + Machine Learning Mastery
  + Towards Data Science Blogs
  + Investopedia
  + Built in Data Science Blogs
  + Scikit- Learn Org