# Mobile Price Range Prediction

Janhavi Shembade, Data Science Trainee @almabetter

# ABSTRACT:

During the purchase of mobile phones, various features like memory, display, battery, camera, etc., are considered. People fail to make correct decisions, due to the non-availability of necessary resources to cross-validate the price. To address this issue, a machine learning model is developed using the data related to the key features of the mobile phone. The developed model is then used to predict the price range of the new mobile phone.

Logistic Regression are used to train the model and predict the output as low, medium, high or very high. The dataset used in this study is taken from the almabetter platform. In order to improve the classification accuracy, Chi-Squared based feature selection method is used. Among 21 features available in the dataset, only top 10 features namely RAM, pixel height, battery power, pixel width, mobile weight, internal memory, screen width, talk time, front camera and screen height are selected and used to train the model. Before applying feature selection, the accuracy obtained using Logistic Regression is 94% and 95% respectively. After feature selection, the accuracy of xgboost improved to 99% and 98% respectively. From the experiments conducted, it is found that XGBoost gave superior performance when compared to other two classifiers.

*Keywords- RAM; Prediction; Machine learning; Artificial neural network*

**INTRODUCTION:**

Mobile nowadays is one of the apps with the most sales and purchases. New mobiles are released each day with new versions and more apps. Hundreds and thousands of cell phones are sold and bought every day. So here the prediction of the mobile price class is a case study for the given type of problem i.e. finding an optimal product. The same work can be done to estimate the real price of all goods, such as vehicles, motorcycles, generators, engines, food items, medication, etc. Several apps are very important for estimating mobile prices, for example Mobile Processor.

In today's busy human life the timing of batteries is also very critical. Mobile size and thickness are also important determinants of decision. Internal memory, camera pixels and the consistency of the video must be remembered. Internet surfing is also one of the most significant limitations of this 21st-century technological period. And so is the list of several features dependent on those, it is decided on mobile size. So we're going to use all of the above features to determine whether the smart-phone will be very-economical, economical, expensive or very costly.

**BUSINESS UNDERSTANDING:**

Mobile nowadays is one of the most selling and purchasing devices. Every day new mobiles with new versions and more features are launched. Hundreds and thousands of mobile phones are sold and purchased on a daily basis.In the competitive mobile phone market companies want to understand sales data of mobile phones.and factors which drive the prices.In this problem we don't need to predict the actual price but a price range indicating how the price is.

**PROBLEM STATEMENT:**

Price is the most effective attribute of marketing and business.we need to consider many of the factors to predict the actual price. All the customers are first worried and think “If he would be able to purchase something with given specifications or not”. So to estimate price is the basic purpose of the work.

**The main goal of the project is to:**

This paper explains the prediction of a stock using Machine Learning. The technical and fundamental or the time series analysis is used by most of the stockbrokers while making the stock predictions. The programming language used to predict the stock market using machine learning is Python.

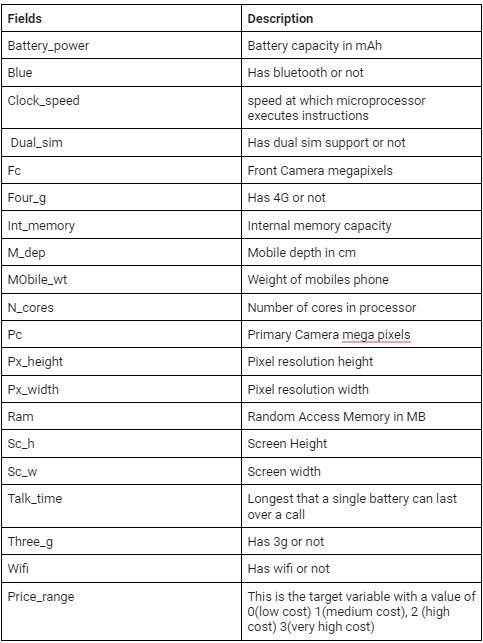
**DATA DESCRIPTION:**

The data description phase starts with an initial data collection and proceeds with activities in order to get familiar with the data. Identifying data quality problems, discovering first insights into the data and detecting interesting subsets to form hypotheses from hidden information are activities of this step. Data which is collected from different mobile price manufacturer companies.It had 2000 rows and 21 columns. Most columns are related to mobile factors. Based on the business understanding of the data 21 columns was chosen to analyse the data

**DATASET PREPARATION:**

The Mobile price range prediction has 2000 rows and 21 columns. The feature “Price Range” is the target variable. Below Table shows the data features.

**Data-set description**



**EXPLORATORY DATA ANALYSIS:**

If we want to explain EDA in simple terms, it means trying to understand the given data much better, so that we can make some sense out of it. It was also used to produce a value distribution and identify missing values, and outliers.

EDA is a process of examining the available dataset to discover patterns, spot anomalies, test hypotheses, and check assumptions using statistical measures. In this chapter, we are going to discuss the steps involved in performing topnotch exploratory data analysis

In statistics, A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modelling or hypothesis testing task.EDA in Python uses data visualization to draw meaningful patterns and insights

* **DATA ANALYSIS:**

This is one of the most crucial steps that deals with descriptive statistics and analysis of the data. The main tasks involve summarizing the data, finding the hidden correlation and relationships among the data, developing predictive models, evaluating the models, and calculating the accuracies. Some of the techniques used for data summarization are summary tables, graphs, descriptive statistics, inferential statistics, correlation statistics, searching, grouping, and mathematical models.

* **DATA CLEANING**

After completing the Data Sourcing, the next step in the process of EDA is Data Cleaning. It is very important to get rid of the irregularities and clean the data after sourcing it into our system.

Irregularities are of different types of data.

* Missing Values:-So In this dataset contains 0 values we need to fill with its mean value
* Incorrect Format:-The “sc\_h” and “sc\_w” contain the cm we need to convert it in inches.
* Incorrect Headers:- All headers are in good format
* **DATA TRANSFORMATION:**

Data transformation is the process of normalizing and aggregating the data to further improve the efficiency and accuracy of data mining.

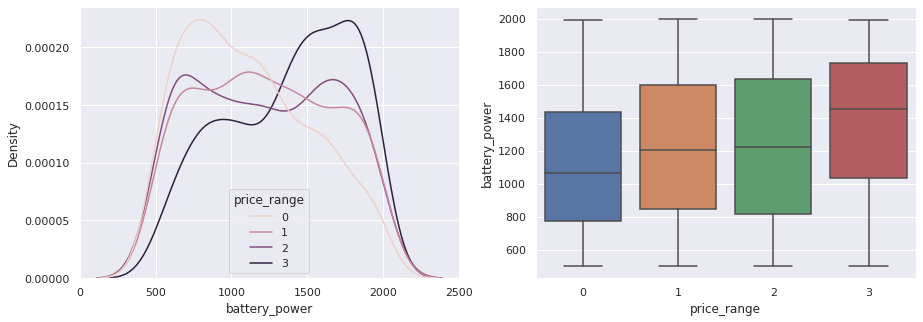
* **DATA DEDUPLICATION:**

It is very likely that your dataset contains duplicate rows. Removing them is essential to enhance the quality of the dataset.

So there is no duplicate value present

* **BIVARIATE ANALYSIS:**

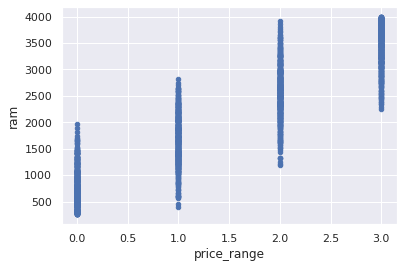
If we analyze data by taking two variables/columns into consideration from a dataset, it is known as Bivariate Analysis.



* **a)Numeric-Numeric Analysis:**

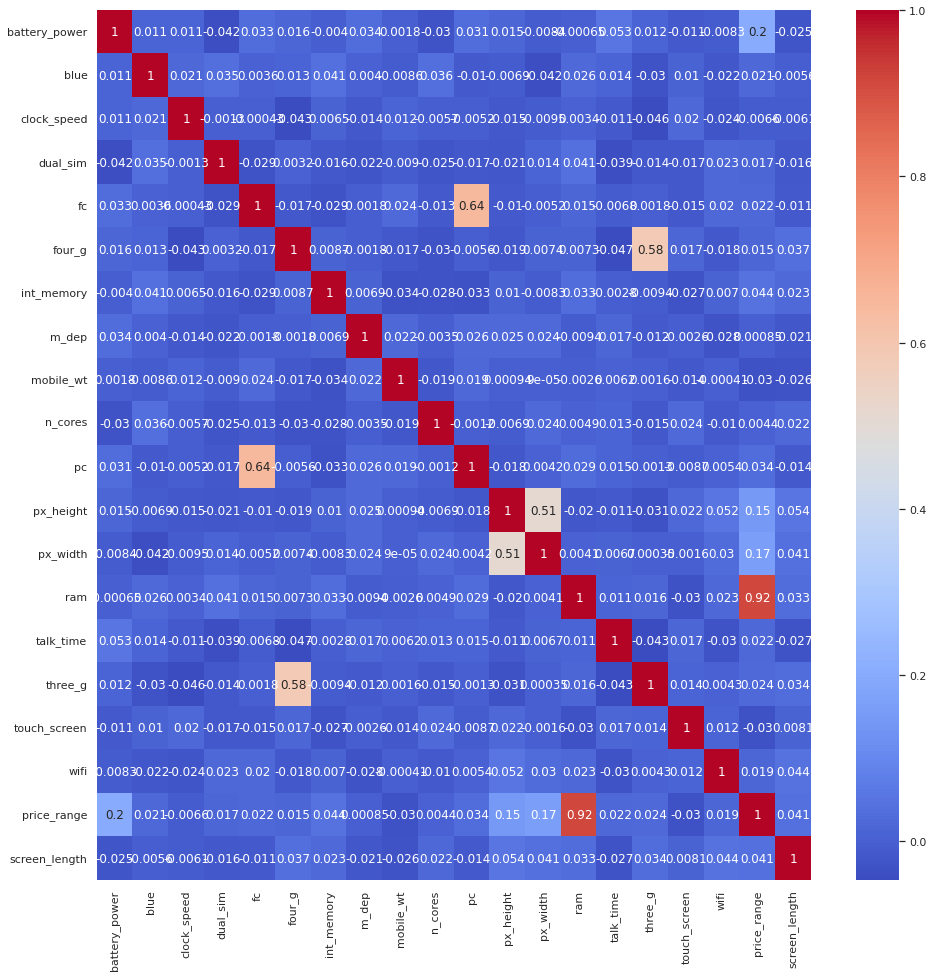
Analyzing the open variables and Date variable from a dataset is known as numeric-numeric analysis. We can analyze it in three different ways.

* Correlation Matrix
* Scatter plot



* **CORRELATION AMONG VARIABLES**:

In words, the statistical technique that examines the relationship and explains whether, and how strongly, pairs of variables are related to one another is known as correlation.Each variable is highly correlated with each other.

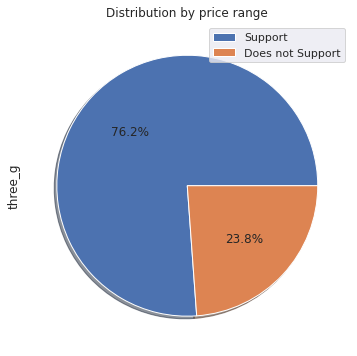


There is some collinearity in feature pairs ('pc', 'fc') and ('px\_width', 'px\_height'). Both correlations are justified since there are good chances that if the front camera of a phone is good, the back camera would also be good.

RAM and price\_range shows high correlation which is a good sign, it signifies that RAM will play a major deciding factor in estimating the price range.

* **GRAPHICAL REPRESENTATION OF THE RESULTS:**

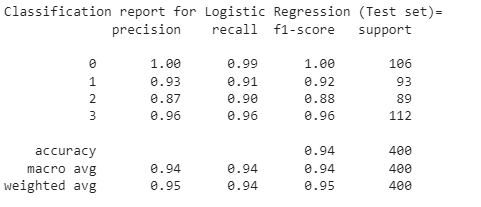
This is an essential step as the result analyzed from the dataset should be interpretable by the business stakeholders, which is one of the major goals of EDA.



So In 3G results we can see that 76.2% of people support 3G networks and 23.8% people does not.

**Logistic Regression:-**

* It is a supervised Ml algorithm used to predict probabilities of an algorithm.It has the target variable i.e discrete/categorical ex:-True/False,yes/no,Approved/decline.
* It uses a linear regression model so we can say logistic regression model.
* It uses log loss function i.e nothing but log loss function.

****

**XGBoost Algorithm:-**

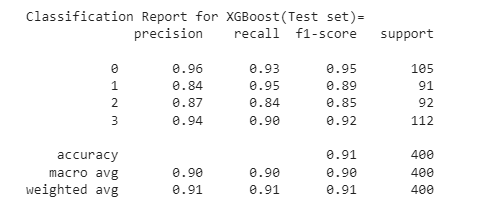
XGBoost is designed for speed, ease of use, and performance on large datasets. It does not require optimization of the parameters or tuning, which means that it can be used immediately after installation without any further configuration.

* XGBoost also has a block structure for parallel learning. It makes it easy to scale up on multicore machines or clusters. It also uses cache awareness, which helps reduce memory usage when training models with large datasets.
* Finally, XGBoost offers out-of-core computing capabilities using disk-based data structures instead of in-memory ones during the computation phase.

XGBoost Train setL-

## 

XGBoost Test Set:-



**CONCLUSIONS:**

A good product can be recommended to a customer by defining the economic range which can best be achieved by mining and analysis of data. In our use case the price range of a mobile was successfully predicted with a high accuracy by training the model for a dataset of two thousand instances with various attributes using logistic and XG boost algorithm with accuracy of 97%.

## 1.Mobile phones have 4 price in ranges

## 2.For bluetooth half of having devices and half don't have any devices

## 3.Price Range for battery power is in increasing order

## 4.Ram has continuous increase with price range while moving from Low cost to

## Very high cost.

## 5.costly phones are light in weight and having RAM,battery power and pixel

## 6.plays a very significant role for deciding prices.

## 7.form all the above experiments we can conclude that logistic regression and,

## XGboosting with using hyperparameters we got the best results.