**MOBILE PRICE RANGE PREDICTION**

**Swapnil Bhagawat, Sumit Zoting,**

**Bhagyashree Bhoyer**

**Data science trainees,**

**AlmaBetter, Bangalore**

**Abstract:**

Mobile phones have become a necessity for every individual nowadays. People want more features and best specifications in a phone and that too at cheaper prices.

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.

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.

My experiment can help you understand what could be the best performing model for our project.

1. **PROBLEM STATEMENT:**

Mobile phones have become a necessity for every individual nowadays. People want more features and best specifications in a phone and that too at cheaper prices.

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.

**Mobile range prediction csv data:**

The contents present in Dataset are:

**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 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).

1. **INTRODUCTION:**

Exploratory Data Analysis (EDA) is for understanding and analyzing the datasets given to get better insights into the data and can be used for taking important business decisions. We can summarize the main characteristics of the datasets and can plot the data visually by using Python libraries. We can store our data in data frames by using the Pandas library from Python.

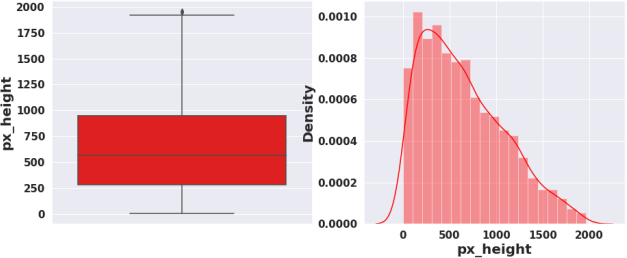
1. **STEPS INVOLVED:**
   1. **Data Exploration:**

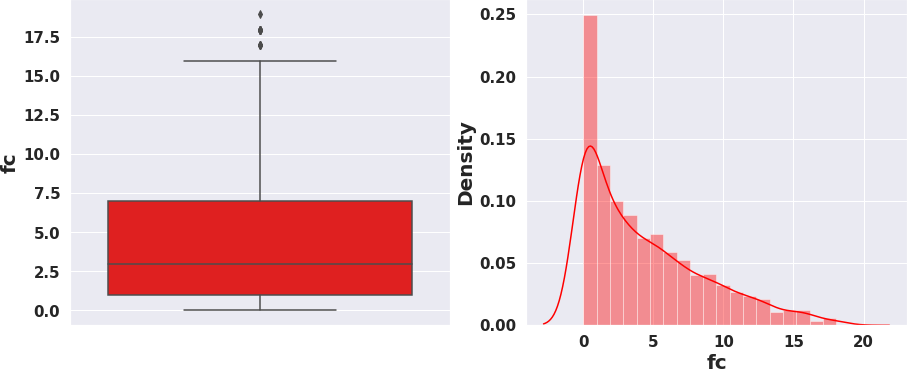
There are data sets which were given mobile-price range prediction csv dataset in which one dataset was having 2000 rows and 21features.

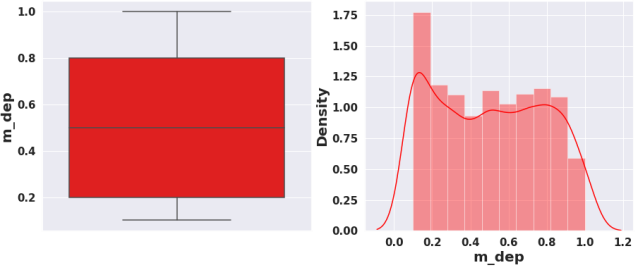
* 1. **Handling Mismatch values in data.**

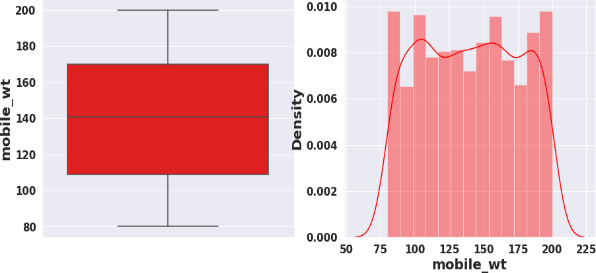
Missing values are imputed using the K-Nearest Neighbors approach where a Euclidean distance is used to find the nearest neighbors.

* Zero Missing values after handling mismatch from the data**.**
* 0 duplicates**.**
* **Checking outliers and distribution of numerical variables:**

****

****

****

****

* 1. **Data Analysis:**

It is a systematic way of describing and illustrating, condensing, and evaluating data by systematically using statistical or logical methods.

* 1. **Data Visualization:**

Python's libraries provide plenty of features with which users can create highly customized, elegant, and interactive plots, making data visualization with Python one of the most widely used features in today's data science environment.

Data Visualization libraries in Python are:

* **Matplotlib:** With a Python library, you can visualize arrays in 2D. NumPy is used to create Matplotlib, which is written in Python. There is a wide variety of plots in Matplotlib, including line, bar, scatter, histogram, etc. that can help us understand trends, patterns, correlations.
* **Seaborn:** A Python library for representing statistics with datasets, Seaborn is a dataset-oriented library. This library is built on matplotlib and creates a variety of visual representations. Pandas data structures are incorporated into it. To create informative visuals, the library performs mapping and aggregation internally

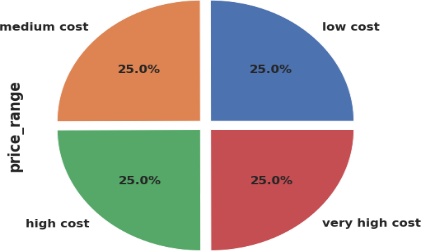
1. **Exploratory Data Analysis:**

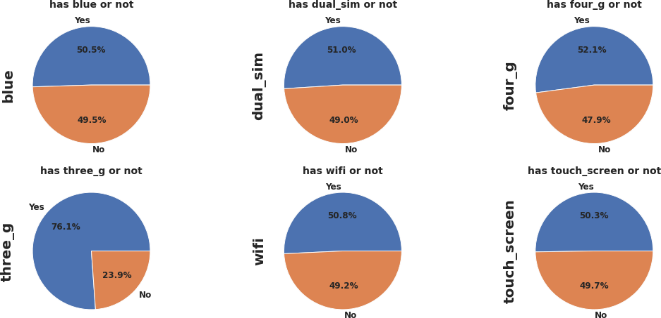
* **Univariate Data analysis**

Our target variable has equal number of observations in each category. Target variable is equally distributed.

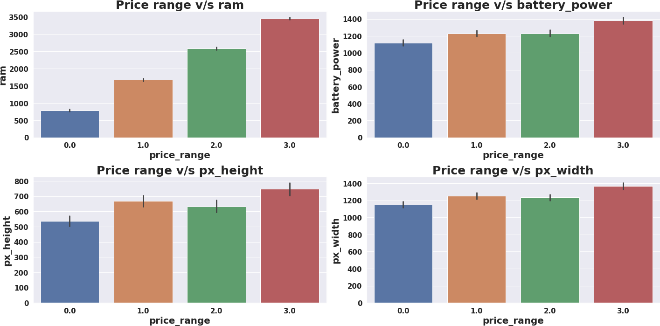
Percentage Distribution of Mobiles having bluetooth, dual sim, 4G,wifi and touch screen are almost 50 %.

Very few mobiles(23.8%) do not have 3G .

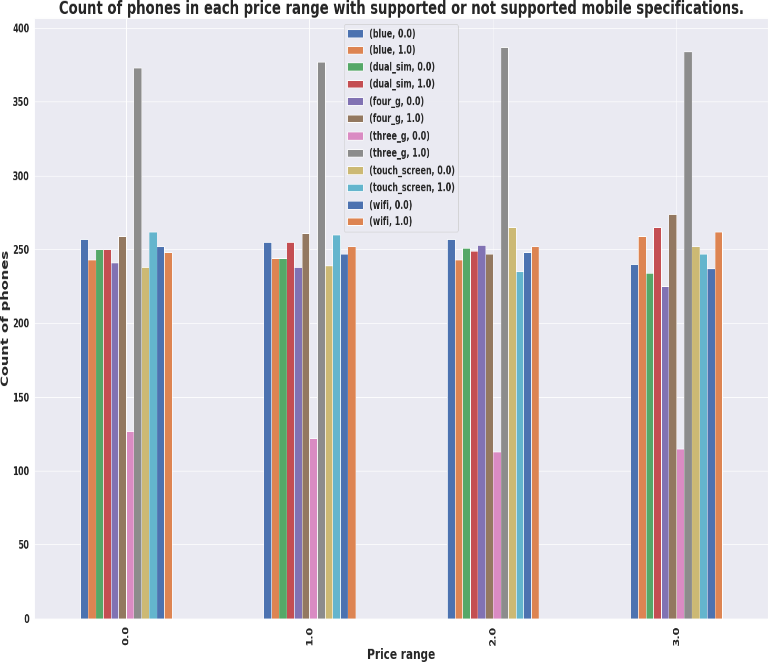


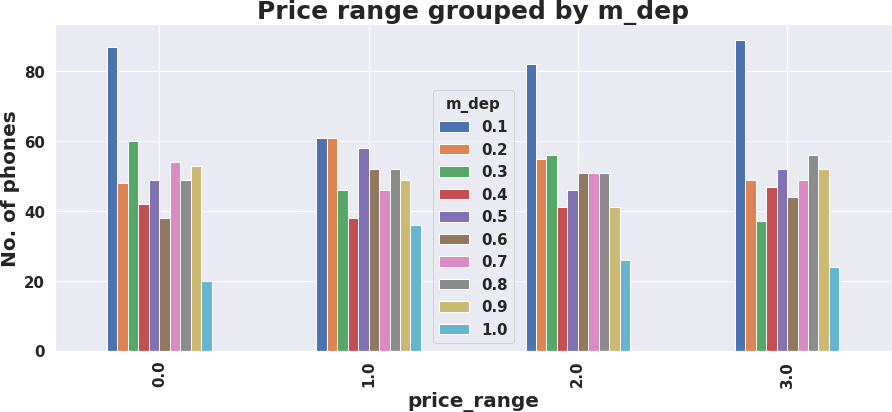


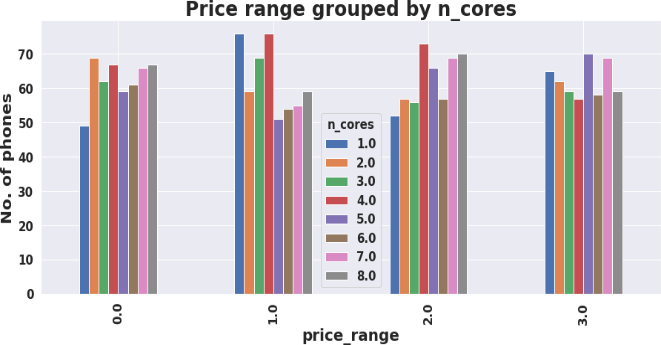
* 1. **Bivariate And Multivariate Analysis**
* Mobiles having RAM more than 3000MB falls under Very high cost category. As RAM increases price range also increases.
* Mobiles having RAM less than 1000 MB falls under low cost category.
* Mobiles with battery power more than 1300 mAh has very high cost. And Mobiles with battery power between 1200 and 1300 mAh falls under medium and high cost category.
* Mobiles with more than 700 pixel height and width more than 1300 has very high cost.



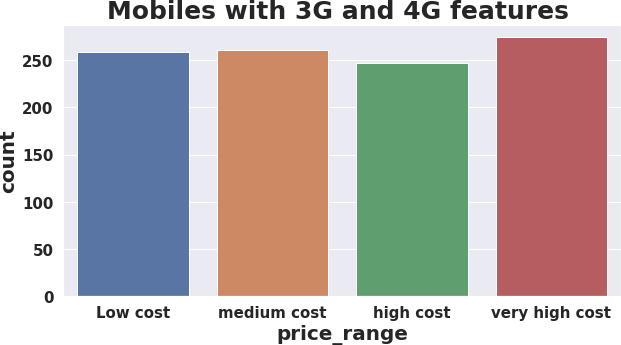
* **Each price range category has equal number of mobiles phones having both supporting and non supporting specifications.**

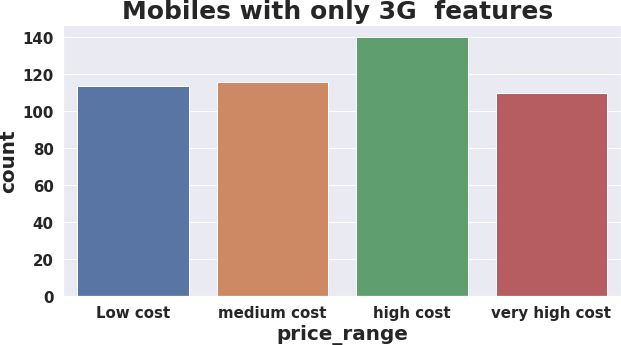






* There are very few mobiles in price range 0 and 1 with lesser no of cores.
* Most of the mobiles in price range 2 and 3 are with high no of cores.
* Number of phones with less thickness is high and count of phones with high thickness is low**.**





* Count of mobiles with 3G and 4G is high in very high cost category.

Count of mobiles with only 3G feature is high in high cost category.

* 1. **Model Selection and Evaluation**

Before building a models we performed the train test split. We kept 25% of the data for test and remaining 75% of the data for training the model.

We compared 6 algorithms and evaluated them based on the overall accuracy score and the recall of the

individual classes.

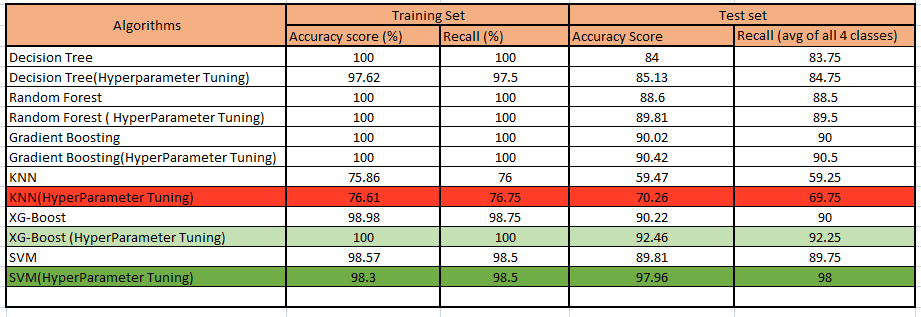
Accuracy is the ratio of the

total number of correct predictions and the total number of predictions.

The recall is the measure of our model correctly identifying True Positives.

1. Decision Tree
2. Random Forest classifier
3. Gradient Boosting Classifier
4. K-nearest Neighbor classifier
5. XG Boost Classifier
6. Support Vector Machine(SVM)

**Evaluation of models:**

****

* Best model came out to be SVM after hyper-parameter tuning.
* XG boost (Hyper-parameter Tuned) can be considered as the second most good model.
* KNN performed very worst.

**Conclusion :**

* We Started with Data understanding, data wrangling, basic EDA where we found the relationships, trends between price range and other independent variables.
* We selected the best features for predictive modeling by using K best feature selection method using Chi square statistic.
* Implemented various classification algorithms, out of which the SVM(Support vector machine) algorithm gave the best performance after hyper-parameter tuning with 98.3% train accuracy and 97 % test accuracy.
* XG boost is the second best good model which gave good performance after hyper-parameter tuning with 100% train accuracy and 92.25% test accuracy score.
* KNN gave very worst model performance.
* We checked for the feature importance's of each model. RAM, Battery Power, Px\_height and px\_width contributed the most while predicting the price range.

**References:**

* https://www.analyticsvidhya.com
* https://www.kaggle.com
* https://www.geeksforgeeks.org/overview-of-data-science/