**Capstone Project**

**MOBILE PRICE PREDICTION**

**Shajad**

**Contents:**

* **Abstract**
* **Introduction**
* **Problem Statement**
* **Steps Involved**
* **Data Visualization**
* **Methods and algorithm that we used**
* **Conclusion**
* **References**

**Abstract:** In the competitive mobile phone market company wants to understand sale data of the mobile phone and factors which drive the prices. The objective is to find out some relation ship between features of mobile phones ( 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.

**Data Description**

Mobile phone price ranges historical data and supplemental information about the stores DatafieldsMost of the fields are self-explanatory. these points are:-

* **Battery power** 🡪 battery power in mAh
* **Blue** 🡪 Has Bluetooth or not
* **Clock speed** 🡪 Speed at which microprocessor execute instruction
* **Dual sim** 🡪 Has dual sim or not
* **Fc** 🡪 front camera megapixels
* **Pc** 🡪 Primary camera pixels
* **Four g** 🡪 has 4g or not
* **Int memory** 🡪 internal memory capacity
* **M deep** 🡪 mobile depth in cm
* **Mobile weight** 🡪 weight of mobile phones
* **N cores** 🡪 no. of cores in processor
* **Px width** 🡪 pixel resolution width
* **Px height** 🡪 pixel resolution height
* **Ram** 🡪 random access memory in mb
* **Sc w** 🡪 screen width
* **Talk time** 🡪 Longest that a single battery can last over a call
* **Three g** 🡪 has 3g or not
* **Wifi** 🡪 has wifi or not
* **Price range** 🡪 This is the target variables with a value of (0=low cost),(1=medium cost),(2=heigh cost),(3=very heigh cost)

**Introduction:**

we are provided with historical sale and features data of the previously sale. The task is to forecast the "price range" column for the test set .

**Problem Statement:**

This data set contains information of mobile phones that information includes such as fc, pc, sc width, bluetooth, price range, battery power, dual sim, wifi, three g, four g, etc . All personally identifying information has been removed from the data.

Explore and analyze the data to discover important factors that affect the price range and sales od the mobile phone

We will tackle the problem statement in the following steps:

Step 1: Data Overview.

Step 2:Data Preparation and Cleaning.

Step 3: Visualizing the Univariate and Bivariate features.

Step 4: Correlation Analysis.

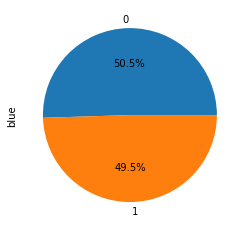
Step 5: Concluding Analysis.

**Steps Involved:**

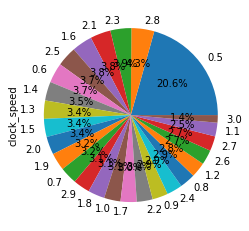
* **Importing Packages:** Importing the various libraries that will help us analyze our dataset properly with visual graphs.
* **Data Overview :** We load the data and go over the basic features, shape and datatypes of various variables.
* **Data Preparation and Cleaning:** We use various features of python to create combined features of date and time and other variables which can be simplified. We also drop columns that we used to creat the new columns. Finally we take care of outliers.
* **Univariate and Bivariate Analysis:** We use seaborn and matplotlib to plot graphs starting with one variable graphs and then plotting two or more variable graphs to understand the variables and it spread and range with its frequency
* **Concluding Analysis:** Here after plotting and analyzing all the data we finally make predictions and remarks about our analyzes.

**Data Visualization:**

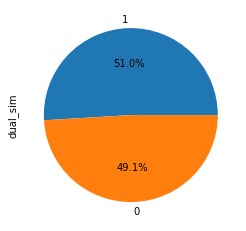
Lets understand various relation among target and other variables



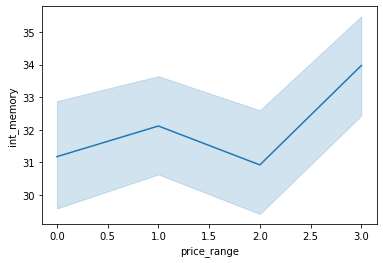
In this pie plot we see that half of the phone don’t have the bluetooth there are 50.5% phones have bluetooth and 49.5% phones are bluetooth less.



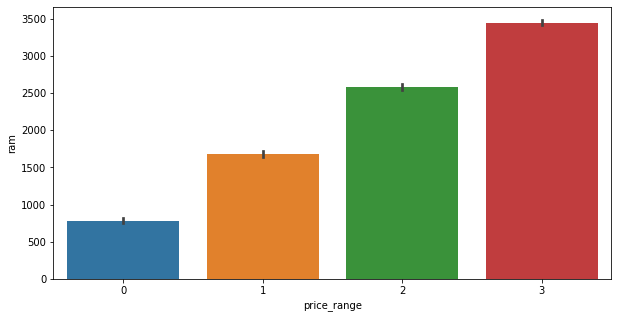
In this pie chart we shoe the major phones have the 0.5 clock speed , it means that the 0.5 clock speed are mostly used in the phones.



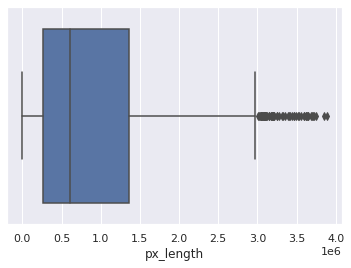
according to this chart 51% phone have dual sim support and 49% phones are single sim supporter



In this chart we can show price range 1 and there



in this chart we show that price range and ram have linear corrilation ram will increase if we increase the ram



In this box plot we see that some outlier are finds after the 3.0 px length.

**Conclusion:**

We perform the some Data wareglings with data set and do the graphical presentation with the data set to find out the relationship between the columns or features.

We create and delete some columns in this data set for make the data more clear and easy to understand.

There are half phones have bluetooth and most usable clock speed is 0.5, 20% phone are using the 0.5 clock speed.

Half of the phone are dual sim operator and half of single sim supported. and also we shoe that if we increase the size of battery power it increase the talk time but there are another relation we find them is if the battery power same in touch and keypad pone, but the talk time of touch is less then the keypad cous of touch screen is bigger and use more battery power.

Ram is increase according to the price range price range 3 have more ram then other and price range 2 have less rem then price range 3 and the more ram then price range 1. and so on. and screen size have positive relation with price range. it increasing according to the price range. we found the some outlier also.In first step we upload the data and then we start to find the duplicate values and nulls values and missing values etc. but in this data there is no missing values and duplicate values. After completion of 1st step in 2nd step we pot the data in form of graph pie plot and bar plot and line plot and also box plot and in box plot we find some outlier but these outliers are not affect the result because they are very less in number. After plot the data we plot the heatmap of the data and then we find the relationship between the feature and we did some feature engineering in the data like combine the features and make another feature etc.

After the feature engineering we apply some algorithm, 1) KNN, 2) SVM, 3) decision tree, 4)random forest, 5) bagging, 6)sacking,7)XG booster, 8) Gradient boosting.

After applying these algorithm we select the best 3 algorithm that gives the best accuracy score. And we select the XG boost, random forest, gradient boosting the give the accuracy between the 88- 91 percent. After selecting the algorithm we do some feature engineering and do some cross validation but there is not much change in the result of the cross validation and we select all as a best that give us 91% accuracy score.

**References:**

* GeeksforGeeks
* Stackoverflow
* Almabetter
* Youtube
* Github