Capstone project -3

Mobile price prediction

By- Ganesh walimbe

Problem statement

- 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.

Points to discuss

- ❑ Objectives
- ❑ Introduction to data
- ❑ Data preprocessing
- ❑ Exploratory data analysis
- ❑ Model building
  - Logistic regression
  - Decision tree classifier
  - Random forest classifier
  - Naive bayes classifier
  - K nearest neighbourhood
- ❑ Conclusions

Objectives

- To find the relationship between features of mobile phones and the selling price of the phone
- To build an prediction model to predict the price range of the mobile phone using supervised learning models .

Introduction to data

In our dataset there are 2000 rows and 21 columns.
There are 21 variables present in the dataset and those variables are as follows :

- 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
- screen or not
- Wifi - Has wifi 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 talking
- 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 preprocessing

Missing value detection:

There are zero missing values present in the dataset

```
#cheking if missing values present in data
data.isnull().sum()
```

```
battery_power     0
blue              0
clock_speed       0
dual_sim          0
fc                0
four_g            0
int_memory        0
m_dep             0
mobile_wt         0
n_cores           0
pc                0
px_height         0
px_width          0
ram               0
sc_h              0
sc_w              0
talk_time         0
three_g           0
touch_screen      0
wifi              0
price_range       0
```

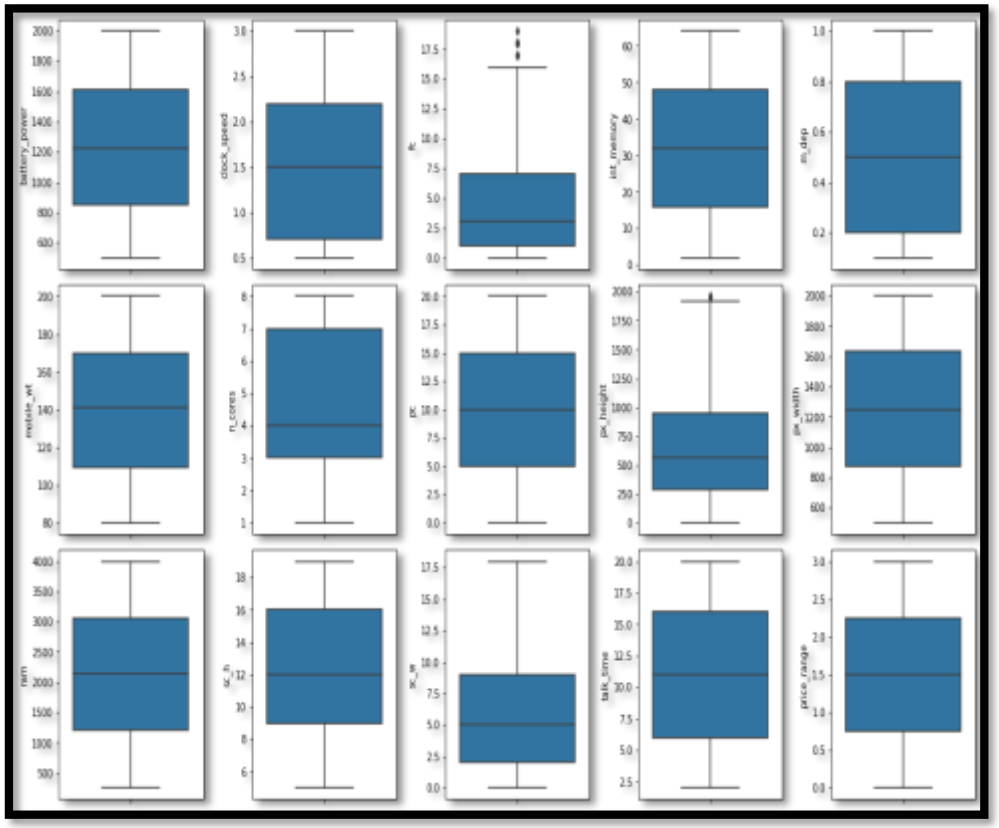
Duplicate values detection:

There were zero duplicate values present in the dataset

```
#cheking if duplicate values present in data
data.duplicated().sum()
```

```
0
```

Data preprocessing



Outliers detection:

From these boxplots we can see that,

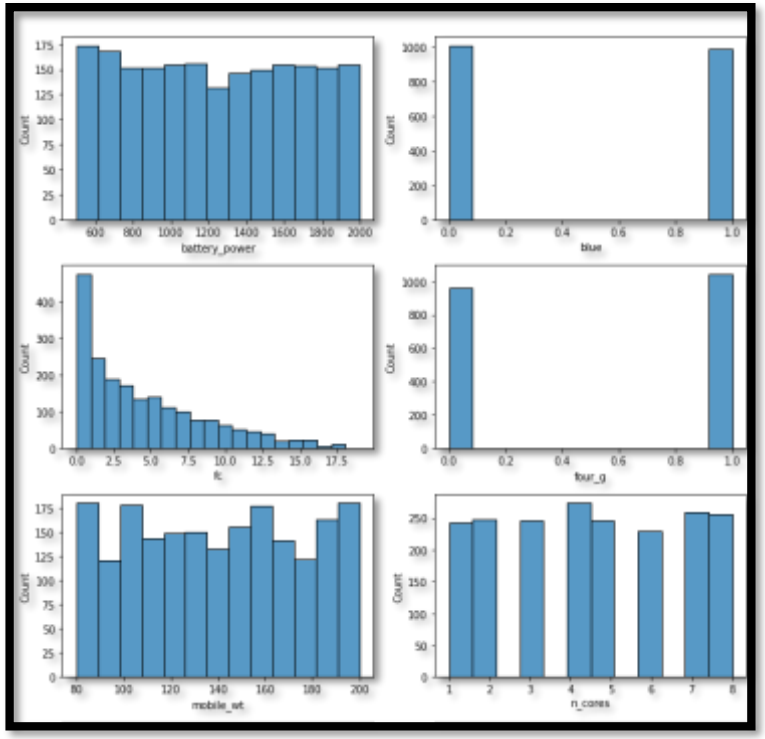
- Almost All the numerical variables does not contains any outliers
- Only fc-front camera megapixel and px_height -pixel height resolution variable have few outliers

Exploratory data analysis

- **Exploratory Data Analysis (EDA)** is an approach to analyze the data using visual techniques. It is used to discover trends, patterns, or to check assumptions with the help of statistical summary and graphical representations.

- **Univariate analysis:**
- Univariate analysis explores each variable in a data set, separately. It looks at the range of values, as well as the central tendency of the values. It describes the pattern of response to the variable. It describes each variable on its own.
- **Bivariate analysis :**
- Bivariate analysis is one of the simplest forms of quantitative analysis. It involves the analysis of two variables, for the purpose of determining the empirical relationship between them.
- **Multivariate analysis:**
- Multivariate analysis is one of the forms of quantitative analysis. It involves the analysis of more than two variables.

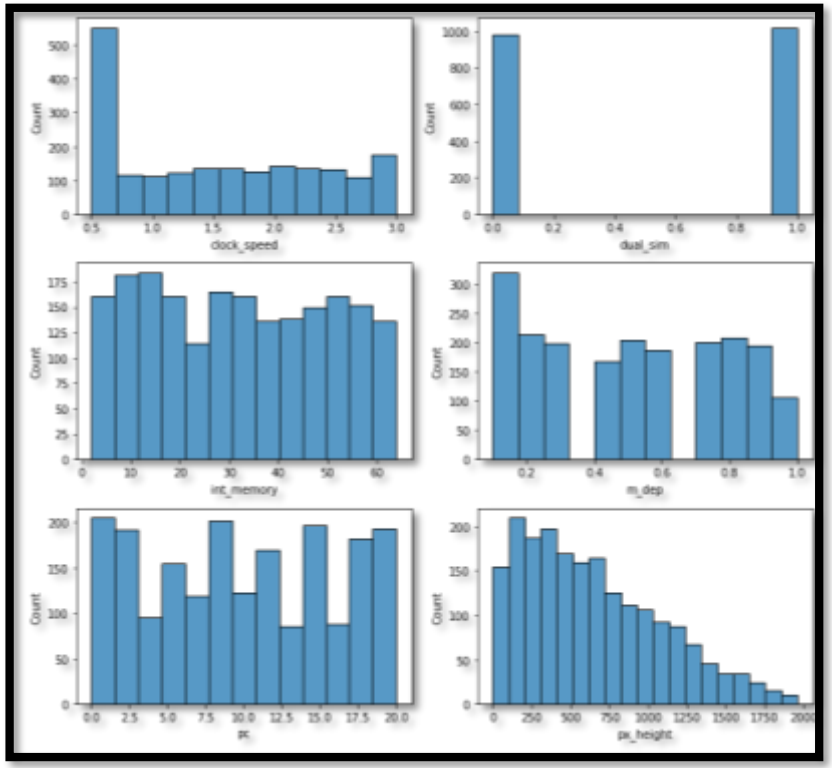
EDA



From theses histogram we can see that

- Mobile phones company is providing all kind kinds of battery power ranges from 600mAH to 2000mAH
- Almost 50% of the phones has Bluetooth and four G network .
- Mobile phone companies are providing more 0 to 8 megapixel front cameras phones as compared to 8 to 18 megapixel front cameras.
- Mobile weights of the phones is ranging from 80gm to 200 gm.
- Mobile companies are providing number of cores of processors from 1 to 8 processors in the phones

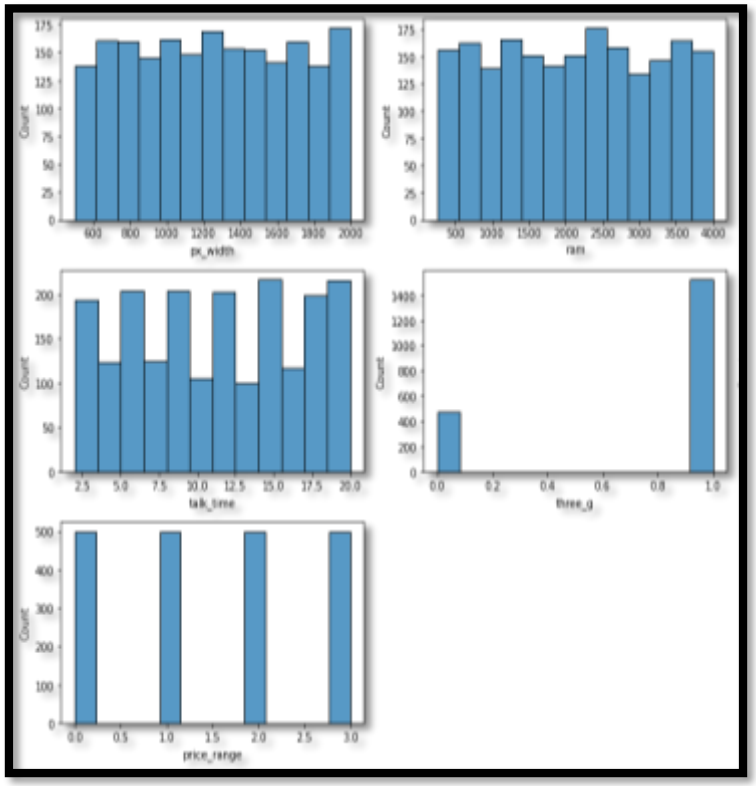
EDA



From histogram we can see that

- The speed at which microprocessor executes instructions for phones is ranges from 0.5 to 3.0
- Almost 50% of the phones has dual sim feature
- Internal memory of the phones ranges from 0 to 60 Gigabytes.
- Primary camera megapixel of the phones varies from 0 to 20 megapixels .
- Mobile Depth in cm is varies from 0.2 cm to 1cm.
- Pixel resolution height of the phones varies from o to 2000.

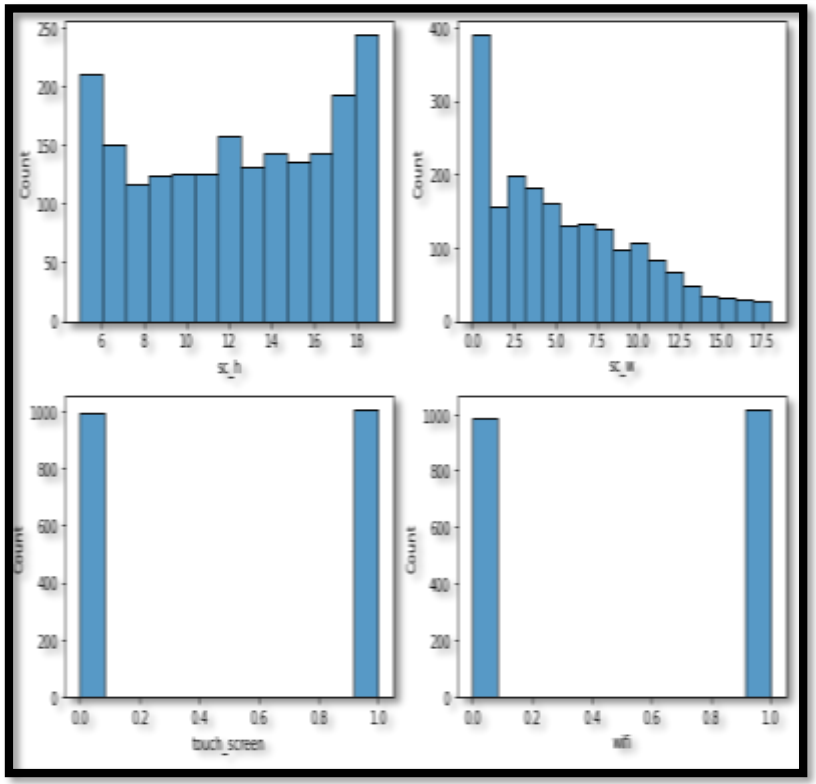
EDA



From histogram we can see that

- Pixel resolution width of the mobile phones are varying from 600 to 2000
- Ram of the mobile phones is varying from 500mb to 4000mb
- Almost 70% mobiles have three G network
- There are four price ranges available for phones

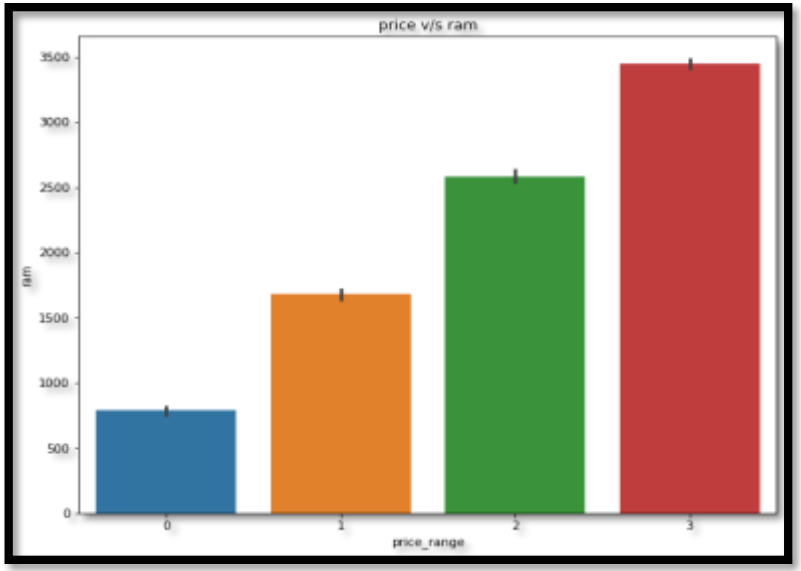
EDA

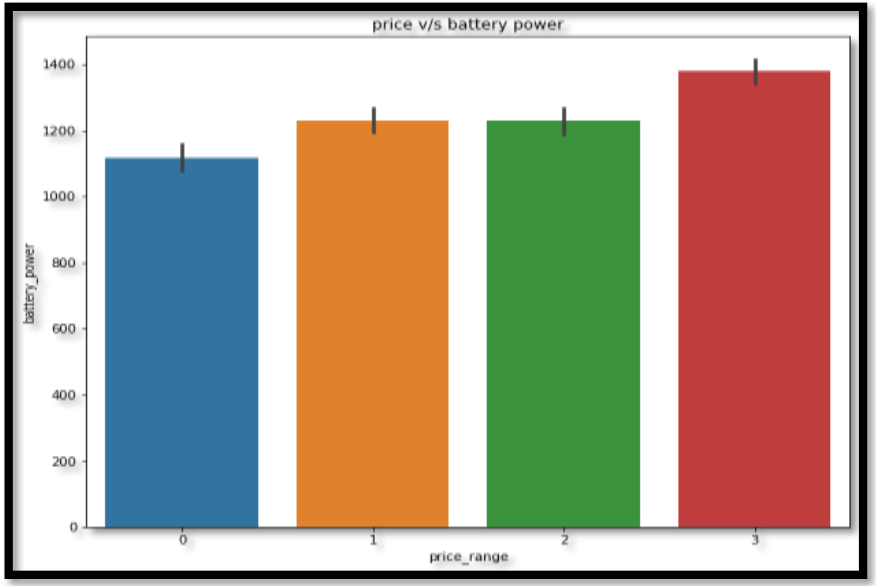


From these histogram we can see that ,

- Screen height of the phones ranges from 6cm to 18cm and Screen width ranges from 1cm to 17.5cm
- Almost 50% of the phones have touch screen
- Almost 50% of the phones have wifi feature .

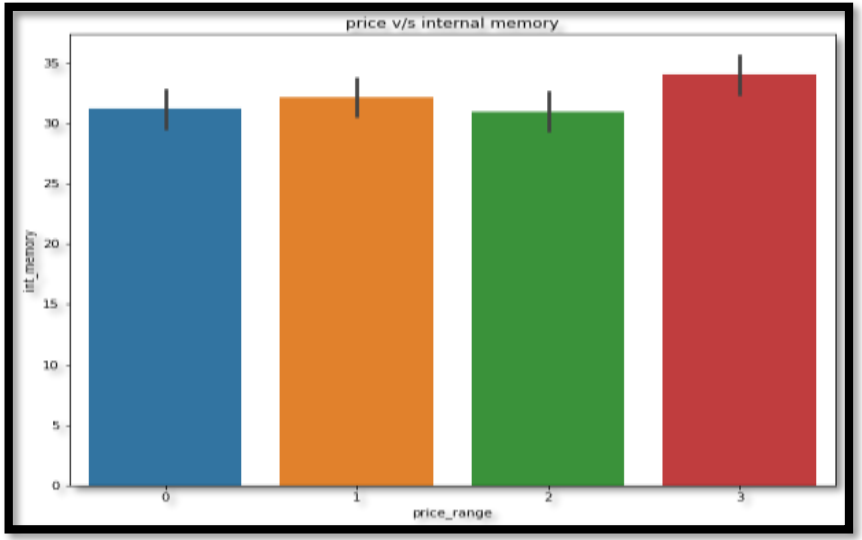
EDA

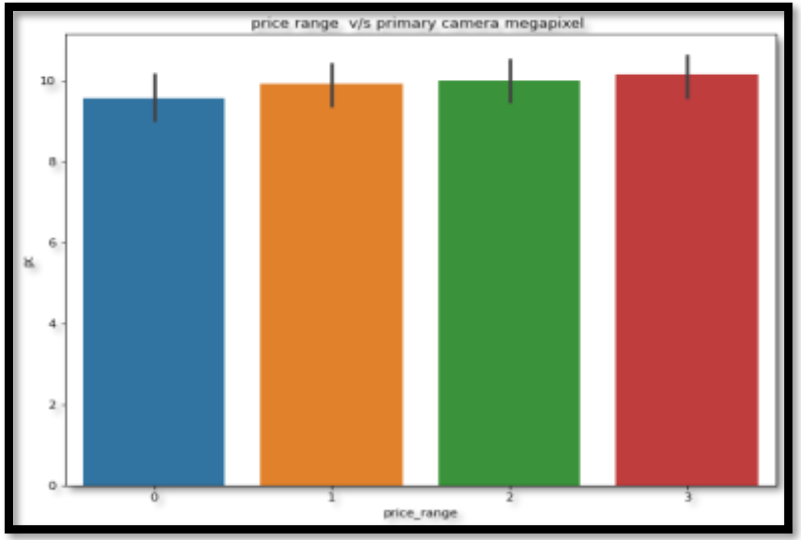




- From price v/s ram barplot we can see that price range is increasing as the ram of the mobile is increasing
- And from price v/s battery power barplot we can see that the price range is almost same for all battery power variants. .

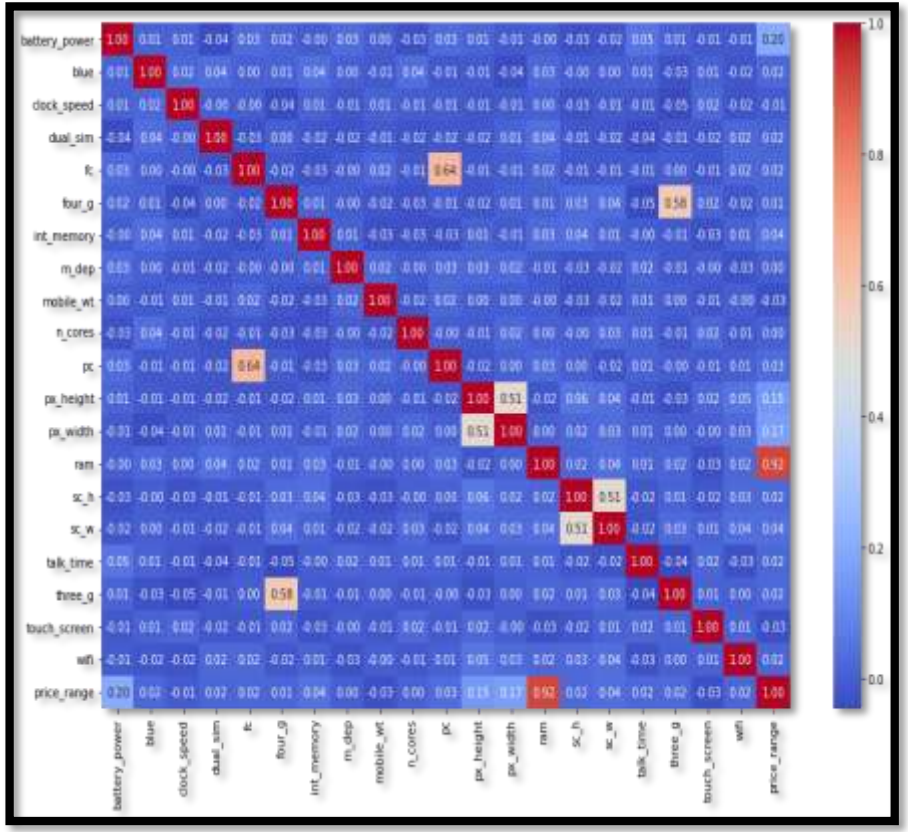
EDA





- from price v/s internal memory barplot we can see that the price range is almost same for all internal memory variants of the mobiles
- And from price v/s primary camera megapixel barplot we can see that the price range is almost same for all primary camera megapixel variants of the mobile.

Correlationplot



Correlation between independent variables :

- From correlationplot we can see that four_g variable is highly correlated with three_g
- Pc is highly correlated with fc variable.
- Sc_h-screen height is highly correlated with sc_w-screen width.
- Px_height-pixel resolution height is also highly correlated with px_width-pixel resolution width.

Correlation between independent variables and dependent variable :

- Ram is highly correlated with our target variable price range
- And battery power is moderately correlated with our target variable price range.

Reducing multicollinearity

- I removed variable three_g and fc because they are highly correlated with with four_g and pc variables respectively
- I also created new variable named px_area using multiplication of px_height and px_width variables.
- I also created new variable named sc_area using multiplication of sc_h and sc_w variables.
- After that i removed px_height,px_width,sc_h,sc_w variables to remove multicollinearity

Building prediction models

To build the prediction we first have to perform training and test on the independent and dependent variables .I am going split data into 80% for training and 20% for testing ,after that i am going to apply various classification models on training and testing data i will find out which model is performing best on the dataset

The models i am going to use are as follows :

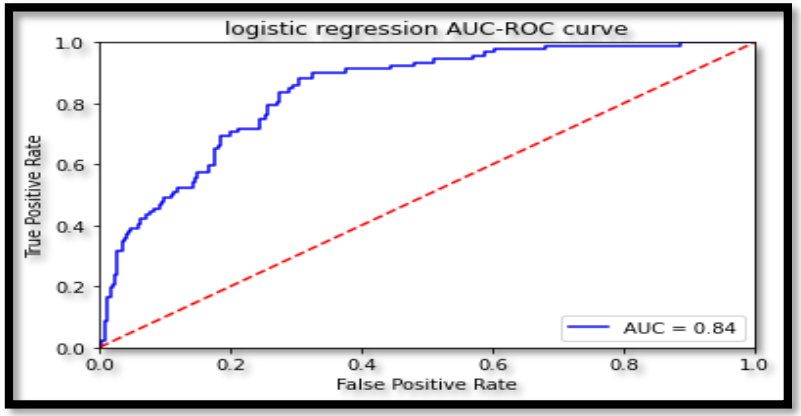
- Logistic regression
- Decision trees classifier
- Random forest classifier
- Naive bayes classifier
- K nearest neighbourhood

Logistic regression

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.

```
#getting training and testing accuracy of logisic regression
print_training_accuracy(train_class_preds,y_train)
#
print_testing_accuracy(test_class_preds,y_test)

Training accuracy is 0.61
Testing accuracy  is 0.61
```



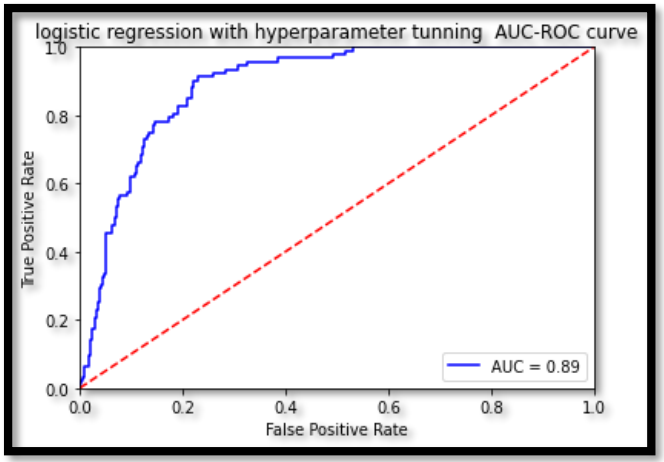
After applying the logistic regression on training and testing data

- The training accuracy for logistic regression is 61%
- The testing accuracy of logistic regression is 61%
- The AUC score of logistic regression is 0.84 .it means that our logistic regression classifier is performing well .

Logistic regression (hyperparameter tunning)

```
#getting training and testing accuracy of decision trees classifier
print_training_accuracy(y_pred_train_lr_ht,y_train)
#
print_testing_accuracy(y_pred_test_lr_ht,y_test)

Training accuracy is 0.685
Testing accuracy  is 0.685
```



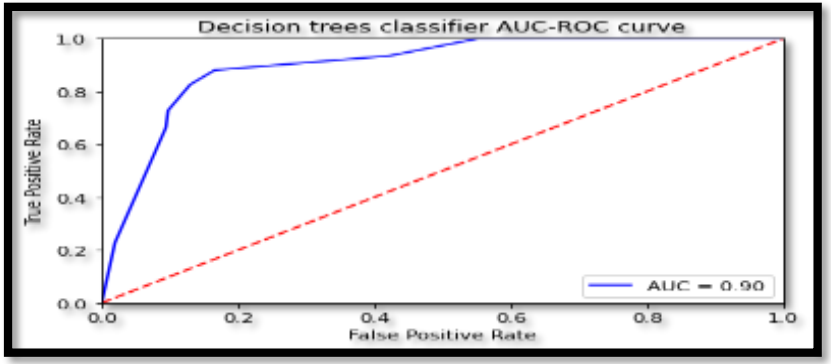
- I Applied hyperparameter tunning to check if accuracy of the model increases.
- After applying hyperparameter tunning the testing accuracy of logistic regression the improved to 68 % .
- And the AUC score of the model has also improved to 0.89 .

Decision trees classifier

Decision Tree is a **Supervised learning technique** that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where **internal nodes represent the features of a dataset, branches represent the decision rules** and **each leaf node represents the outcome.**

```
#getting training and testing accuracy of decision trees classifier
print_training_accuracy(y_pred_train_dt,y_train)
#
print_testing_accuracy(y_predicted_test_dt,y_test)

Training accuracy is 0.796875
Testing accuracy  is 0.78
```



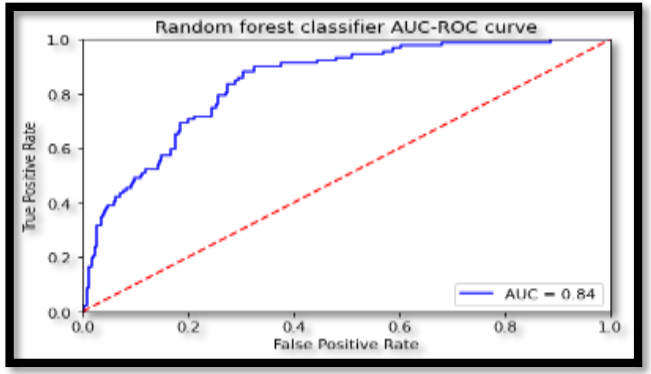
- The training accuracy of decision tree classifier is 79%.
- The decision tree classifier is giving testing accuracy of 78%.
- The AUC score of the decision tree classifier is 0.90. it means that our decision tree classifier is performing well .

Random forest classifier

Random forests or **random decision forests** is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time.

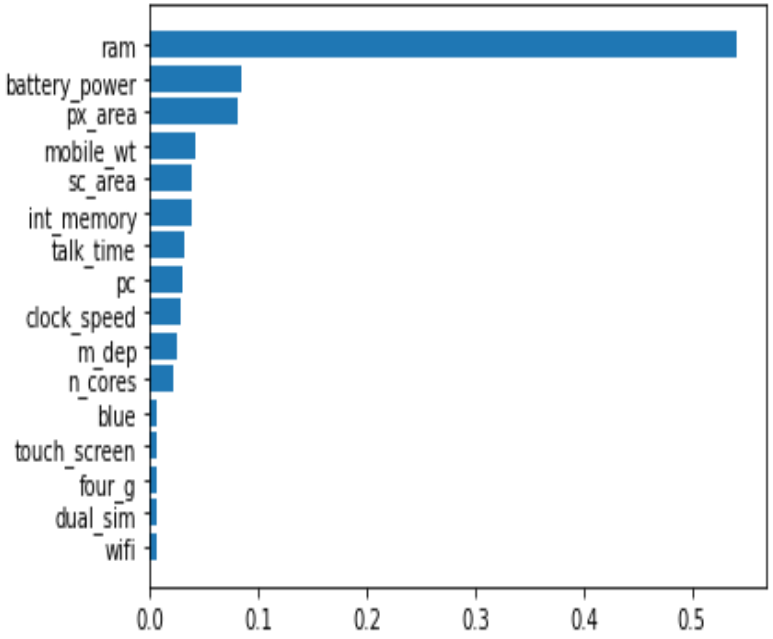
```
#getting training and testing accuracy of random forest classifier
print_training_accuracy(train_preds_rf,y_train)
#
print_testing_accuracy(test_preds_rf,y_test)

Training accuracy is 1.0
Testing accuracy  is 0.8825
```



- The training accuracy of random forest classifier is 100%
- The random forest classifier is giving us testing accuracy of 88%.
- The AUC score of random forest classifier is 0.84. it means that our random forest classifier is performing well .

Random forest classifier



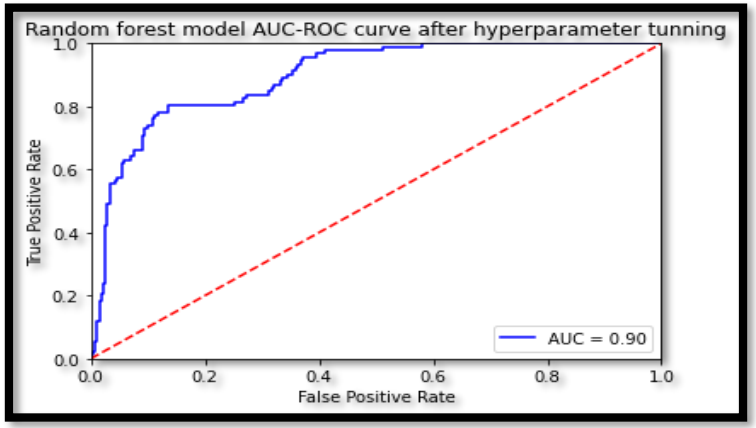
The important features according to random forest model are :

1. RAM–ram of the mobile
2. Battery power
3. Pc_area
4. Mobile_wt
5. Sc_area
6. int_memory

Random forest classifier (hyperparameter tunning)

```
#getting training and testing accuracy of random forest classifier
# after hyperparameter tunning
print_training_accuracy(train_preds_ht_rf,y_train)
#
print_testing_accuracy(test_preds_ht_rf,y_test)

Training accuracy is 0.843125
Testing accuracy  is 0.8075
```



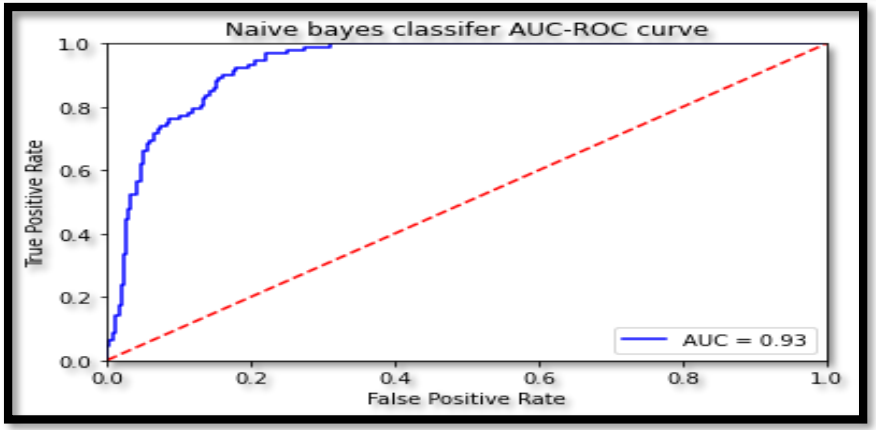
- I applied cross validation and hyperparameter tunning to check if accuracy of the model increases.
- After applying cross validation and hyperparameter tunning the accuracy of random forest classifier the didn’t improved but the AUC score has improved to 0.90 .

Naive bayes classifier

Naïve Bayes algorithm is a supervised learning algorithm, which is based on **Bayes theorem** and used for solving classification problems.

```
#getting training and testing accuracy of naive bayes classifier
print_training_accuracy(train_preds_nb,y_train)
#
print_testing_accuracy(test_preds_nb,y_test)

Training accuracy is 0.796875
Testing accuracy  is 0.8125
```



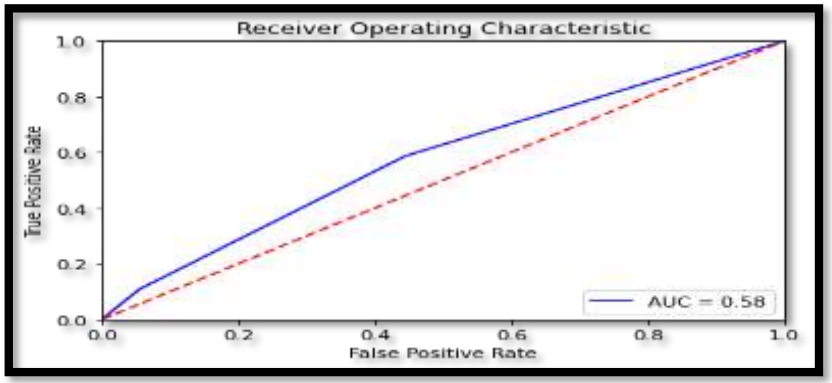
- The training accuracy of naive bayes classifier is 79%
- The Naive bayes classifier is giving accuracy 81%.
- The AUC score of the naive bayes classifier is 0.93. it means that our naive bayes classifier is performing well .

K nearest neighbourhood

K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique . KNN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.

```
#getting training and testing accuracy of k nearest neighbourhood
print_training_accuracy(y_train_pred_kn,y_train)
#
print_testing_accuracy(y_test_pred_kn,y_test)

Training accuracy is 0.700625
Testing accuracy  is 0.3725
```



- The training accuracy of knn is 70%
- The K nearest neighbourhood classifier is giving testing accuracy 33%.
- The AUC score of knn is 0.58.it means model is performing not much well on dataset.

Conclusions

- Conclusions from EDA :
- Price of the mobile phone is directly proportional to the RAM of the mobile .

As the ram of the mobile is increasing the price the mobile is also increasing .

- There are various price ranges available for different types of battery power variants, internal memory variants and primary camera megapixel variants of the mobile phones.
- And from correlation plot we can say that ram is highly correlated with price range of the mobile and battery power also is moderately correlated with price range of the mobile

Conclusions

- Conclusions from prediction models :

index	Model	Training accuracy	Testing accuracy	AUC score
1	Logistic regression	61 %	61 %	0.84
2	Logistic regression (hyperparameter tunning)	68 %	68%	0.89
3	Decision trees	79 %	78 %	0.90
4	Random forest classifier	100 %	88 %	0.84
5	Random forest classifier (hyperparameter tunning)	84 %	80 %	0.90
6	Naive bayes classifier	79 %	81%	0.93
7	K nearest neighbourhood	70 %	33 %	0.58

Conclusions

Therefore from dataframe we can say that random forest model is giving highest accuracy among all the models. so we can use random forest classifier to predict the price range of the mobile phones.

And according to the random forest model the most important factors that are affecting the prices of the mobile phones are: RAM , battery power, pixel area, mobile weight, screen area ,internal memory.

Thank you