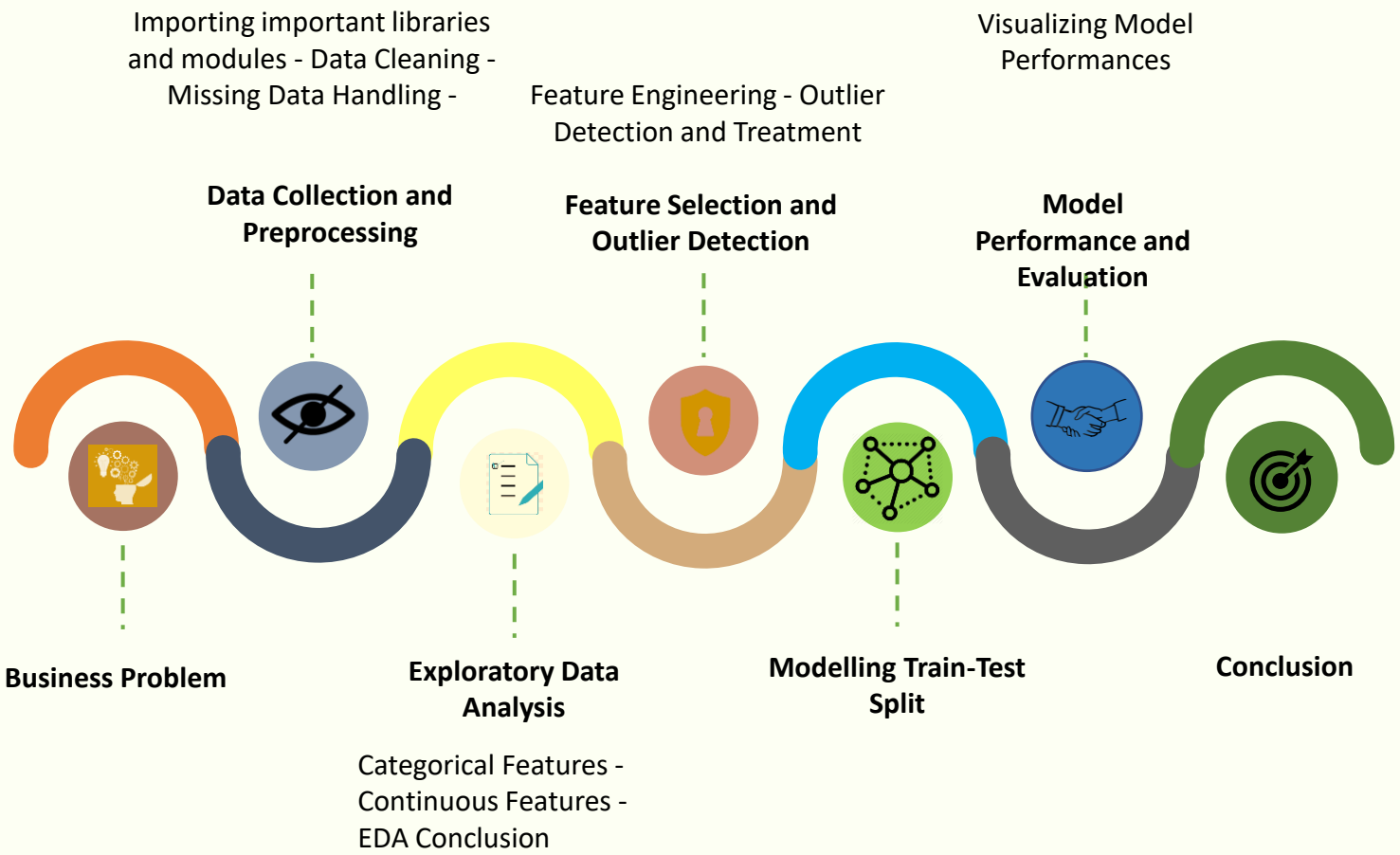
Capstone Project -3

Supervised Machine Learning – Classification

Jouher Lais Khan

Approach


Importing important libraries and modules - Data Cleaning - Missing Data Handling -
Feature Engineering - Outlier Detection and Treatment
Visualizing Model Performances
Data Collection and Preprocessing
Feature Selection and Outlier Detection
Model Performance and Evaluation
Business Problem
Exploratory Data Analysis
Modelling Train-Test Split
Conclusion
Categorical Features - Continuous Features - EDA Conclusion

Problem Description

1. In the competitive mobile phone market companies want to understand sales data of mobile phones and factors which drive the prices.
2. The objective is to find out some relation between features of a mobile phone(eg:- RAM, Internal Memory, etc) and its selling price.
3. In this problem, we do not have to predict the actual price but a price range indicating how high the price is.

Data Description

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

•**Touch_screen -** Has touch screen or not

Wifi - Has wifi or not

•**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

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 Processing

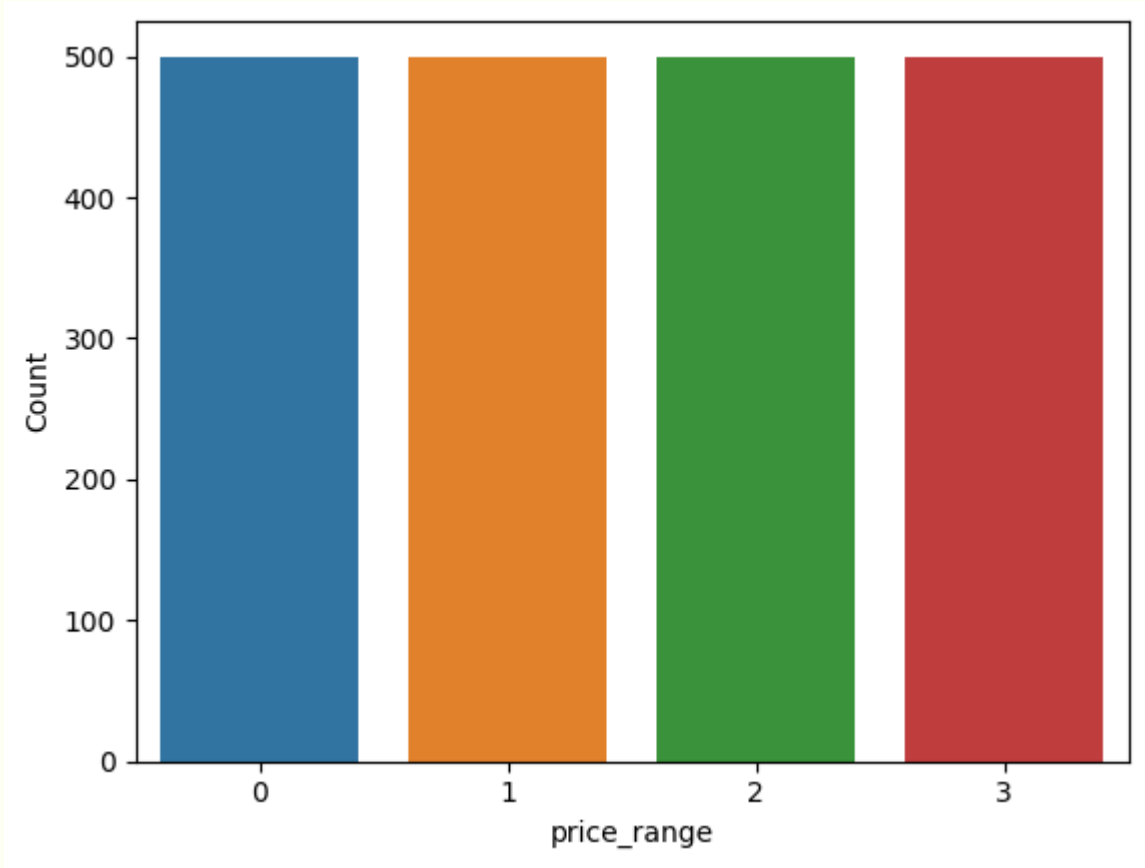
1. Minimum value of px_height and sc_w cannot be zero so we have convert the zero values to the mean value of column.

2. Data Contains no null value

3. Data Contains no duplicates value

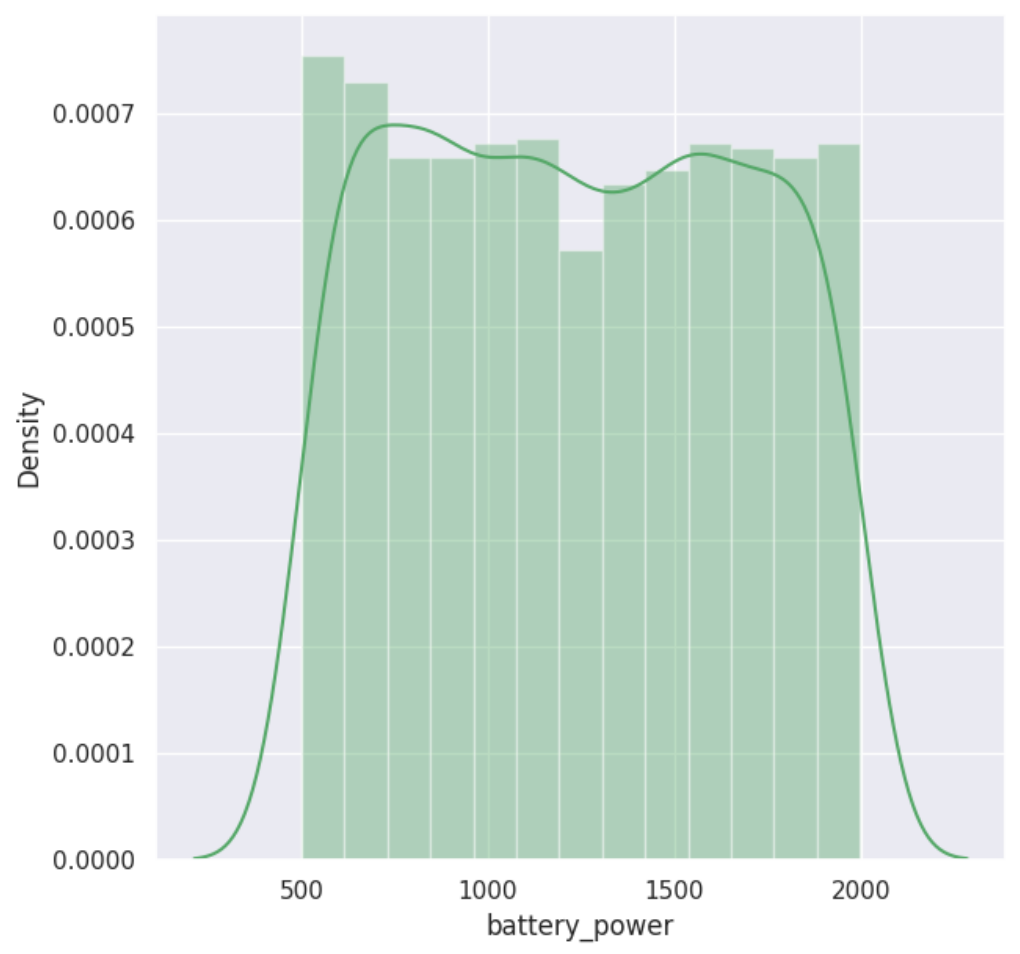
Exploratory Data Analysis

Price



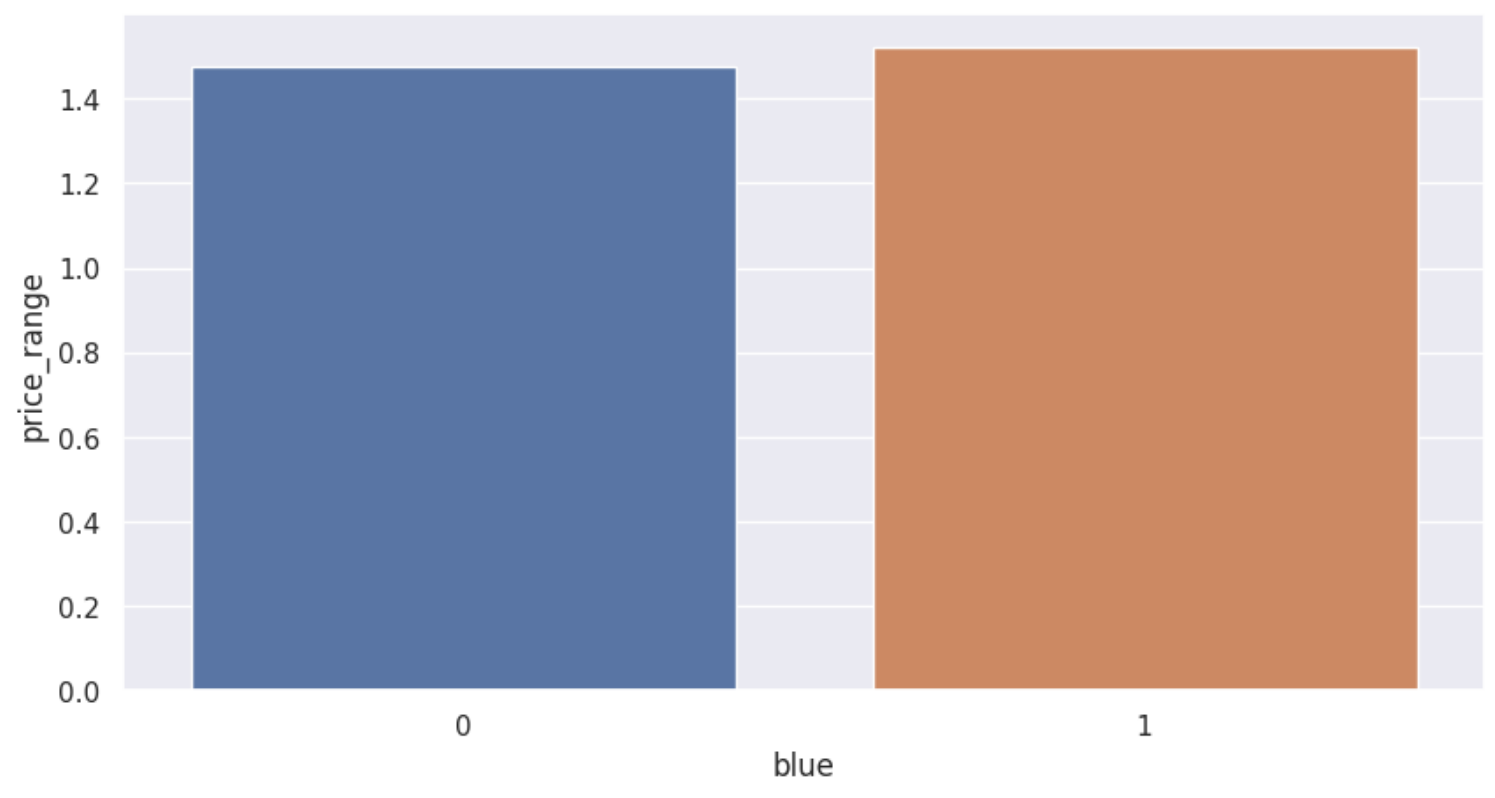
There are mobile phones in 4 price ranges. The number of elements is similar.

Battery Power



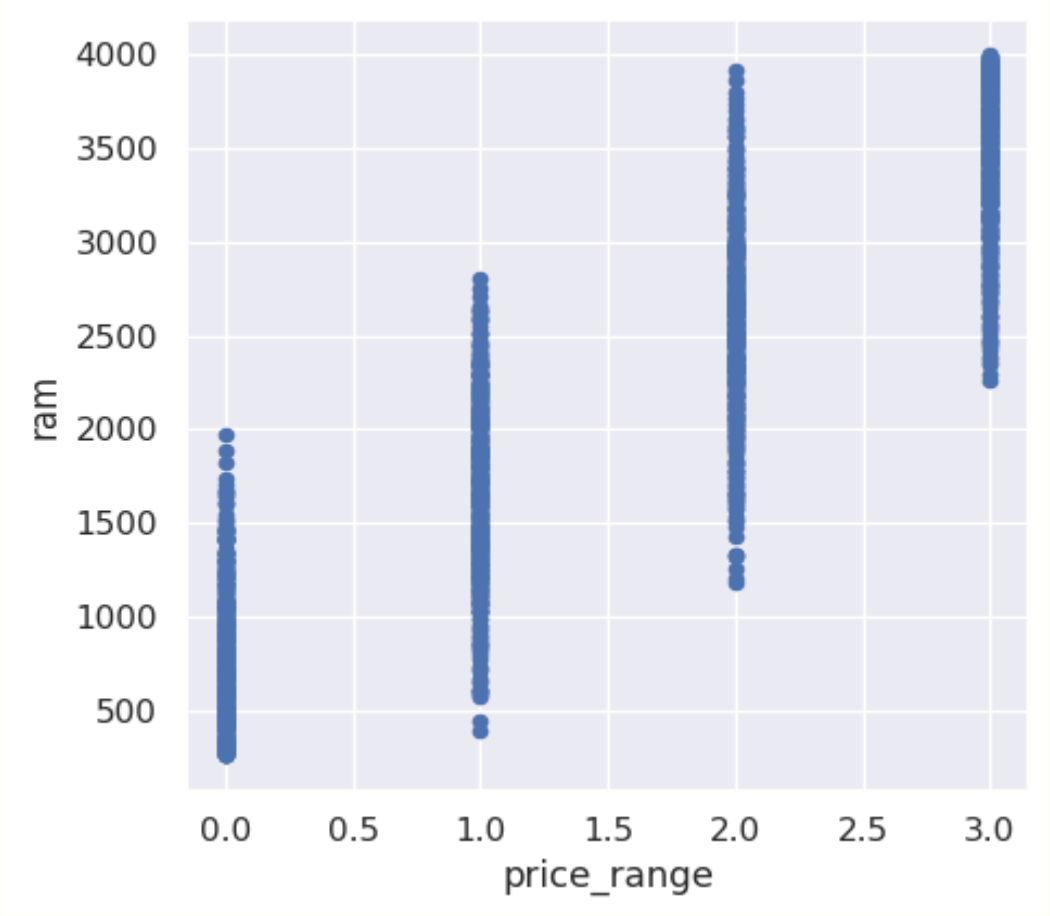
This plot shows how the battery mAh is spread. There is a gradual increase as the price range increases

Bluetooth



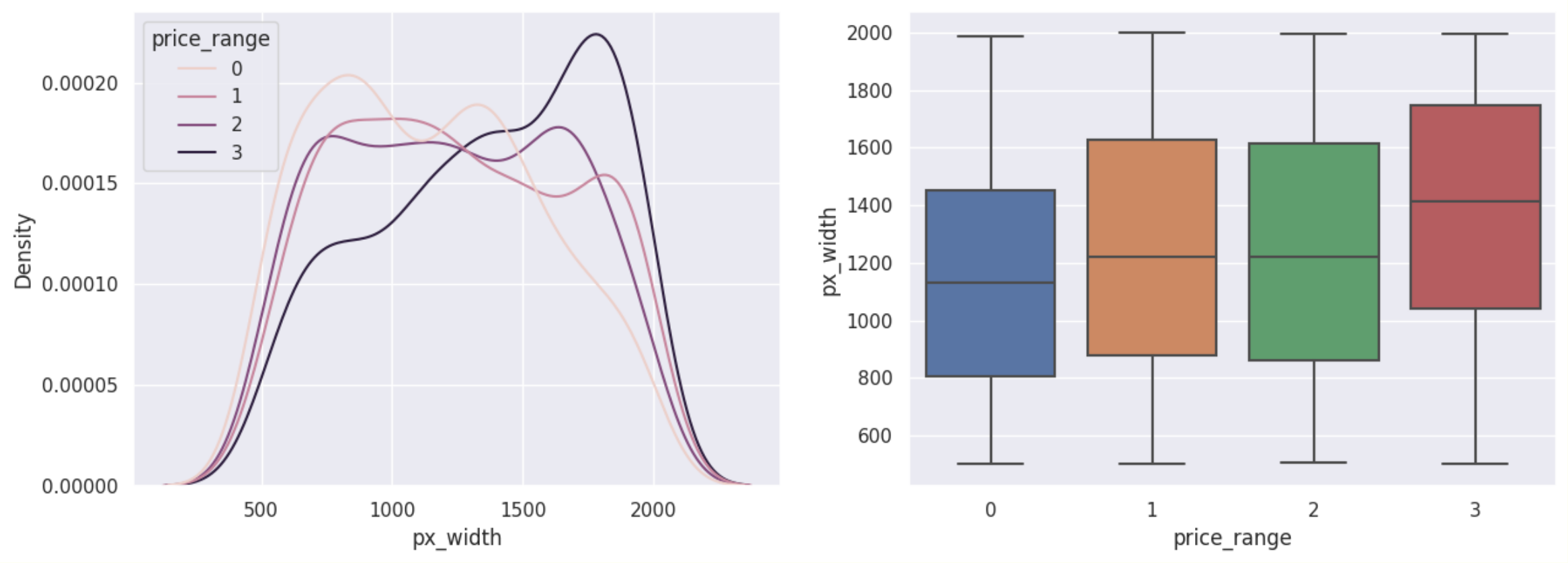
Half the devices have Bluetooth, and half don’t.

Ram



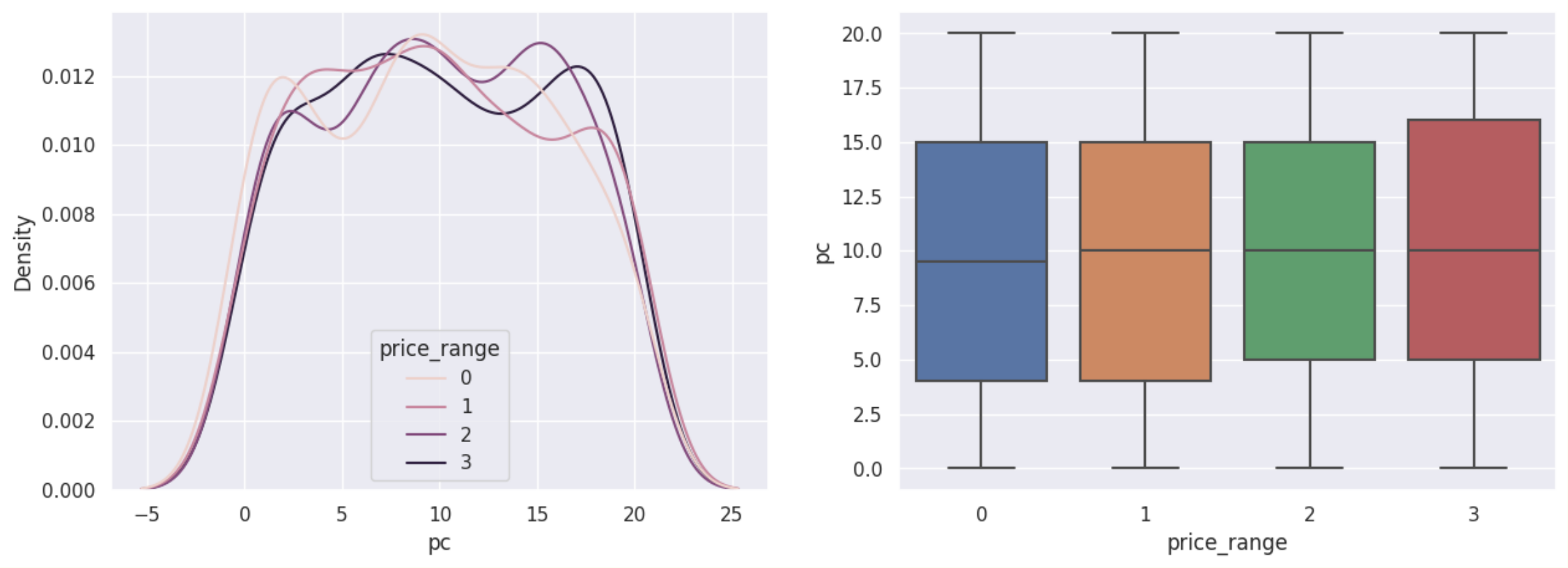
Ram has continuous increase with price range while moving from Low cost to Very high cost

Pixel Width



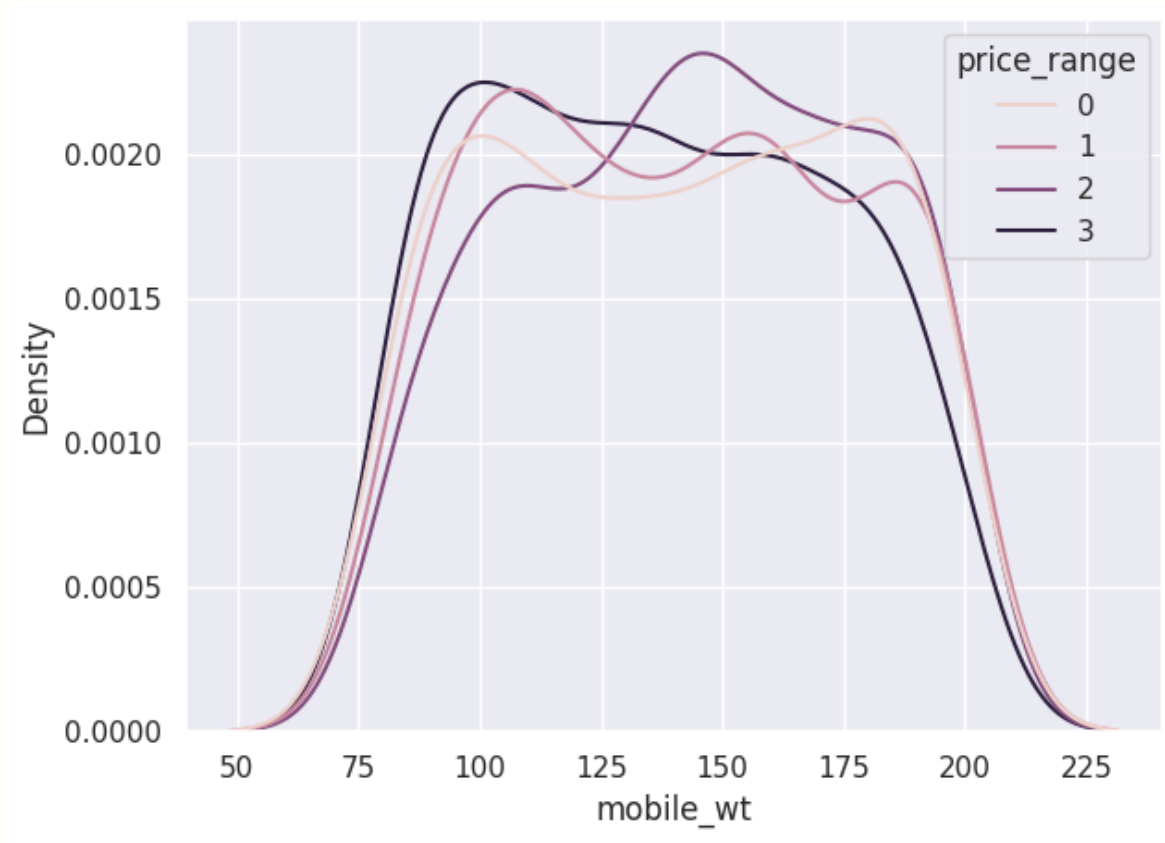
There is not a continuous increase in pixel width as we move from Low cost to Very high cost. Mobiles with 'Medium cost' and 'High cost' has almost equal pixel width. so we can say that it would be a driving factor in deciding price_range.

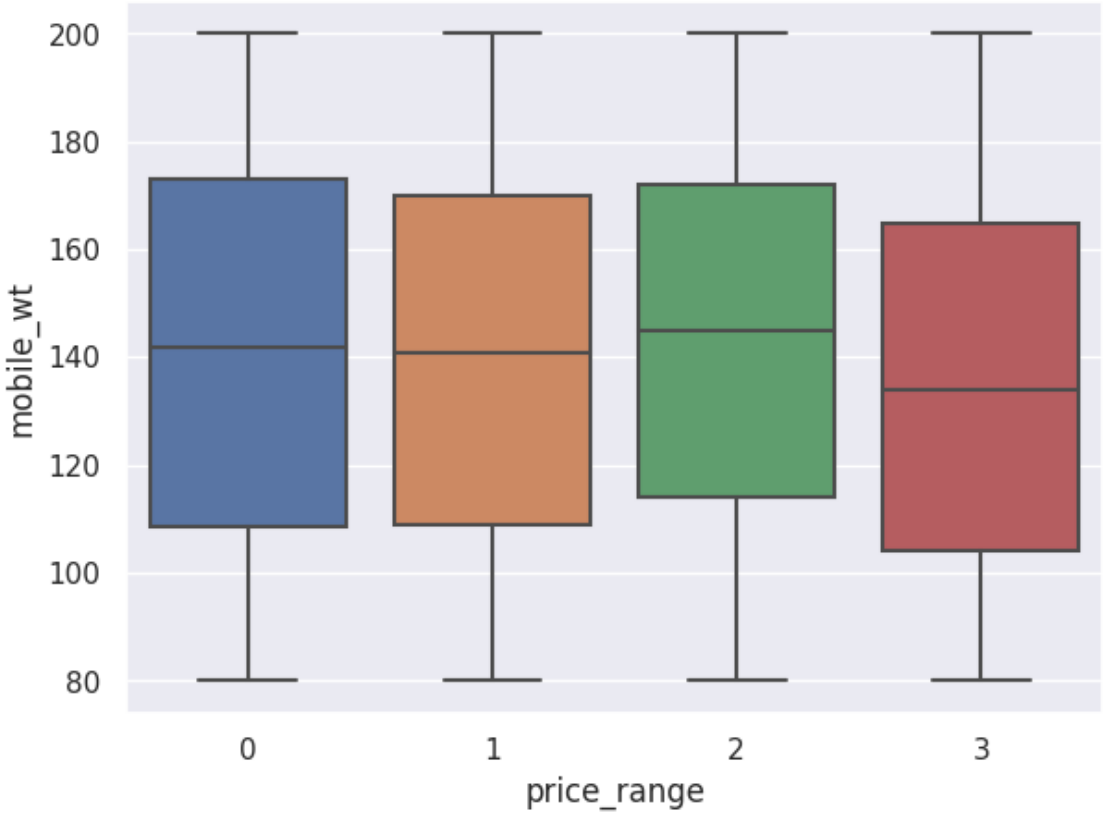
Primary Camera(Mega Pixels)



Primary camera megapixels are showing a little variation along the target categories,this might help in prediction.

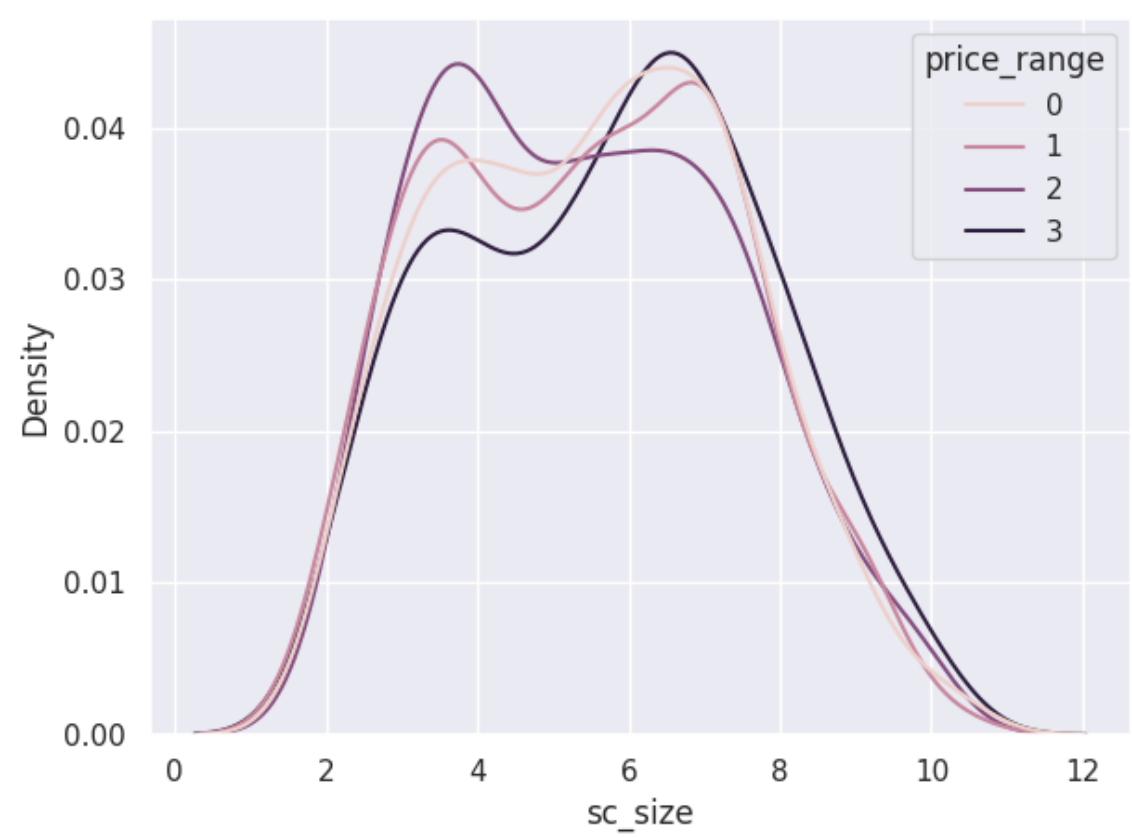
Mobile weight

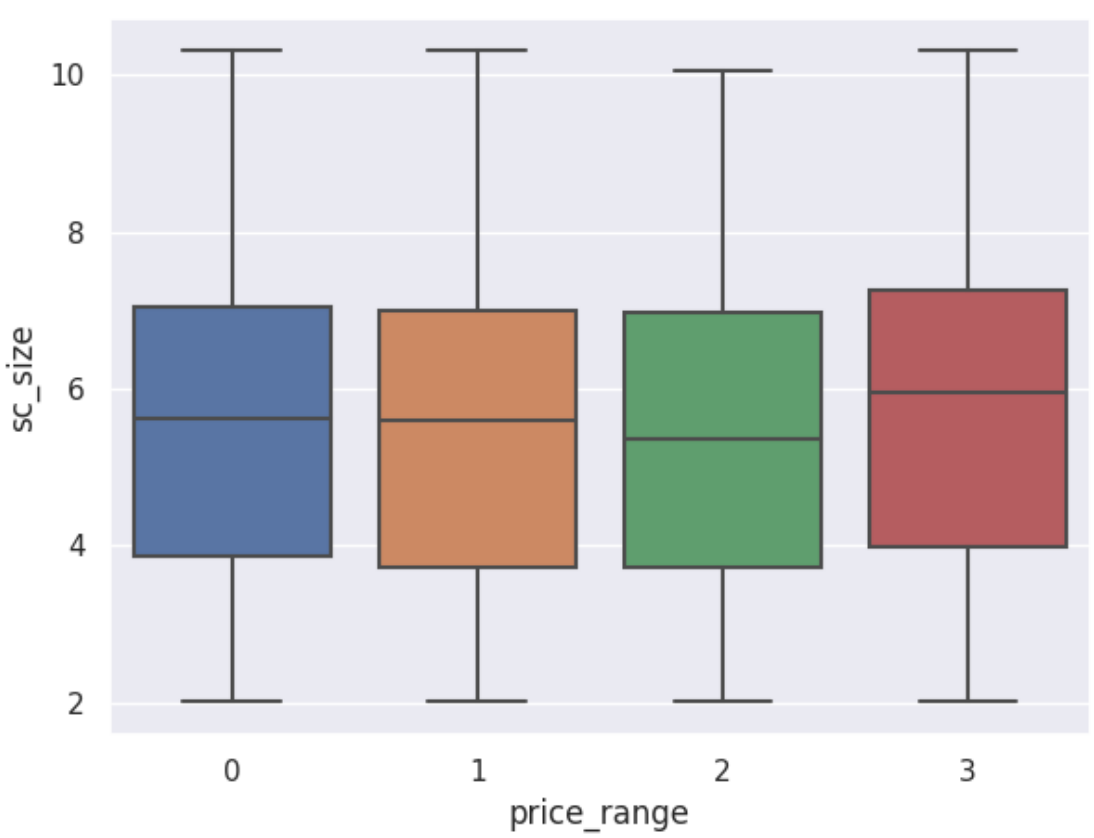




Costly phones are lighter

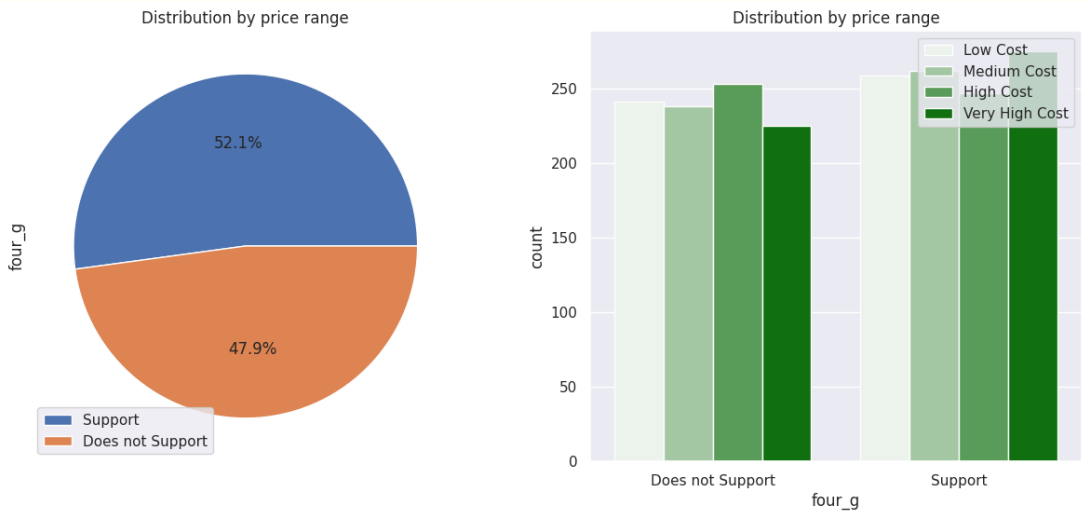
Screen Size

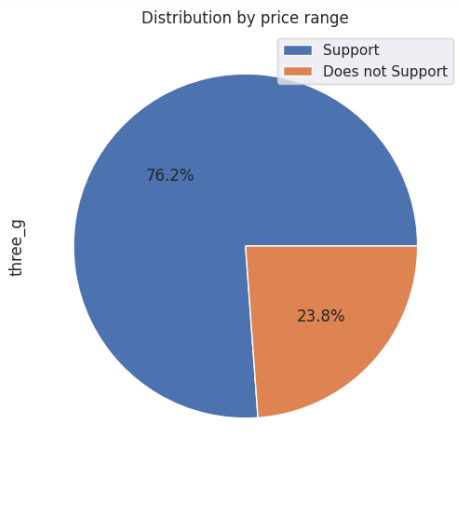




Screen Size shows little variation along the target variables. This can be helpful in predicting the target categories.

3G & 4G





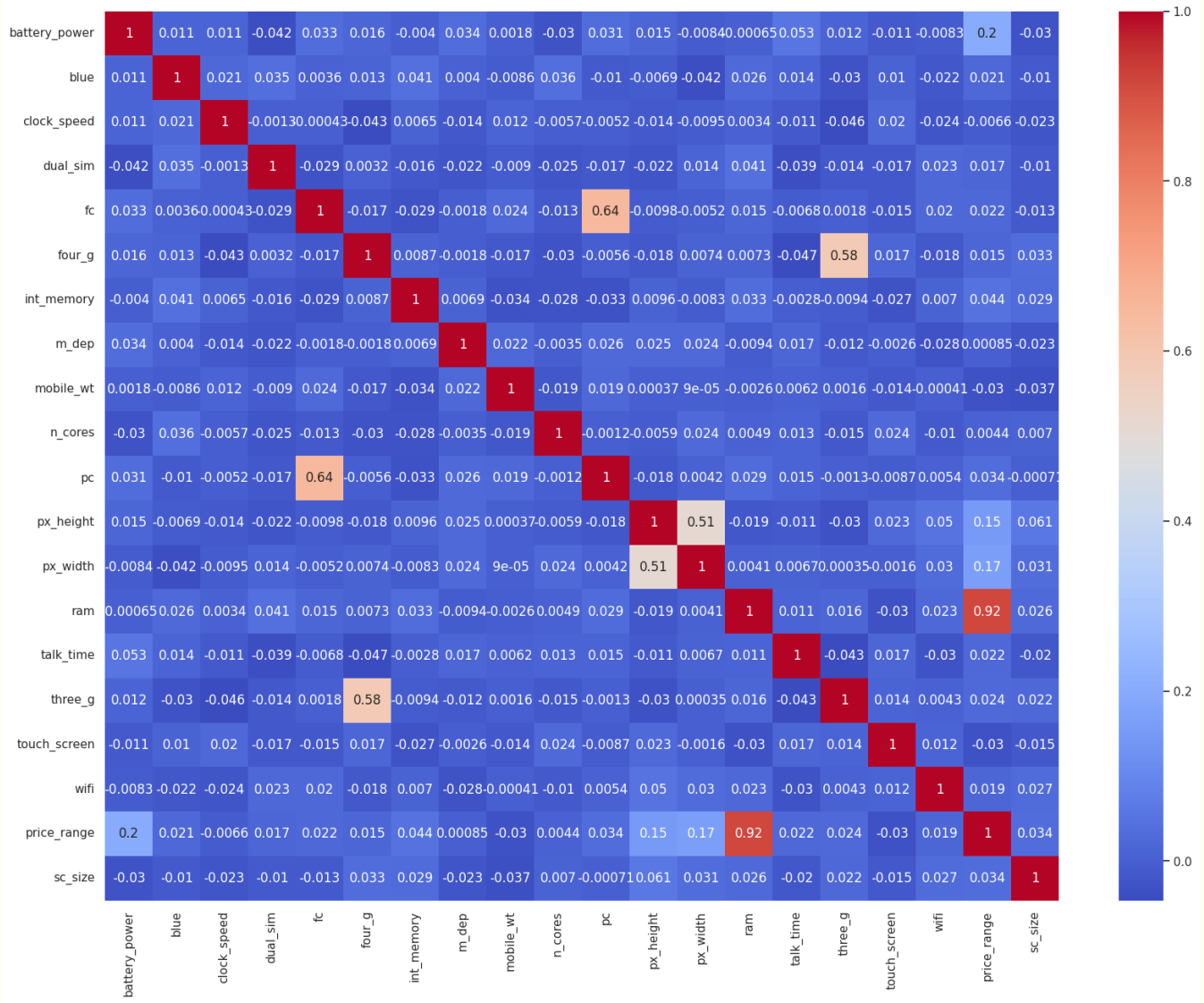
Feature 'three_g' play an important feature in prediction

Correlation Map

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.

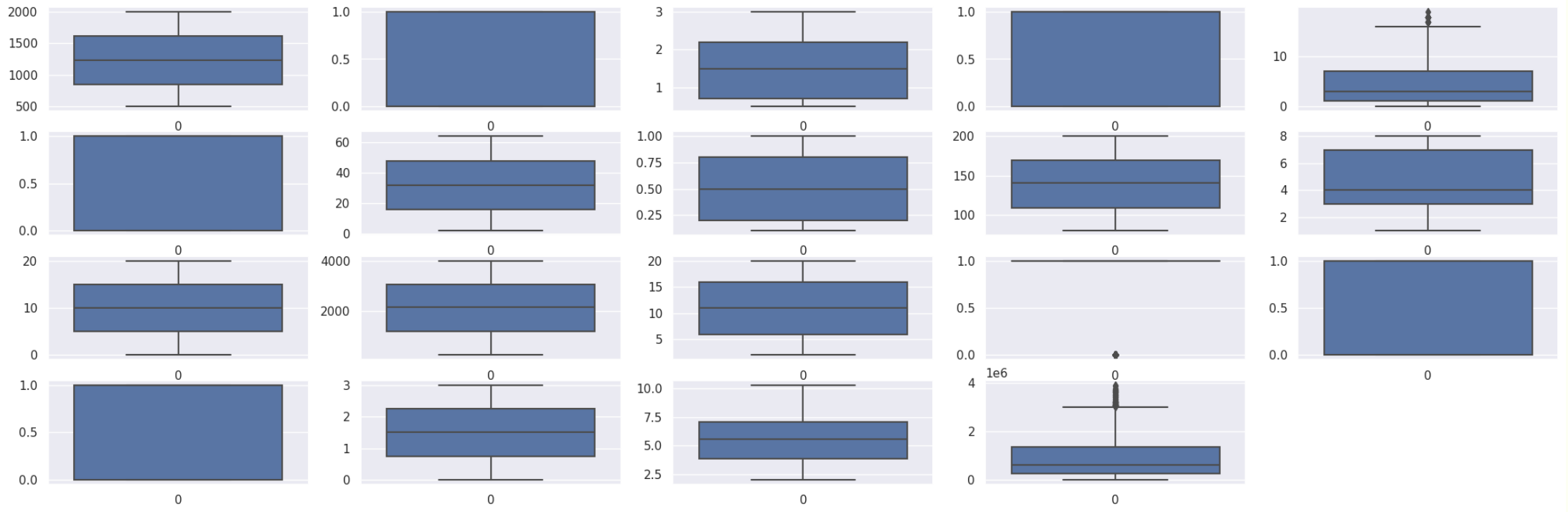
There is some collinearity in feature pairs ('pc', 'fc') and ('px_width', 'px_height'). Both correlations are justified since there are good chances that if front camera of a phone is good, the back camera would also be good.

Also, if px_height increases, pixel width also increases, that means the overall pixels in the screen. We can replace these two features with one feature. Front Camera megapixels and Primary camera megapixels are different entities despite of showing colinearity. So we'll be keeping them as they are.



Outliers Removal

There are almost no outliers in the data



Feature Encoding

1. Creating copy of Data Frame for Modelling.
2. Creating list of final features which will be used in modelling.
3. Creating Sales as dependent variables and features as independent variable.
4. Train-Test Split

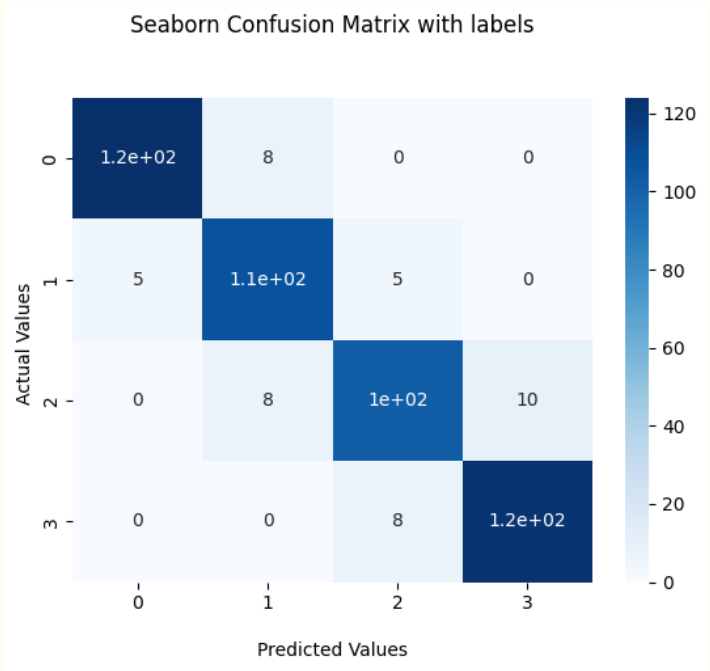
Models Implemented

1. Logistic Regression
2. Decision Tree
3. Random Forest Regression with Hyperparameter tuning
4. xgBoost with Hyperparameter Tuning
5. KNN classifier
6. Naïve Bayes
7. Support Vector Machine

Logistic Regression

Classification report for Logistic Regression (Test set)

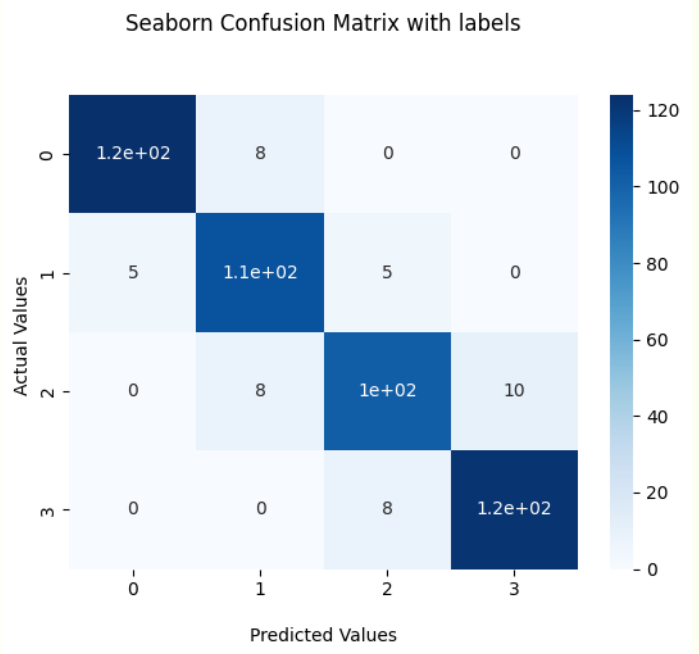
	Precision	Recall	F1-Score	Support
0	0.94	0.96	0.95	129
1	0.92	0.87	0.89	124
2	0.85	0.89	0.87	115
3	0.94	0.92	0.93	132
Accuracy			0.91	500
Macro Avg	0.91	0.91	0.91	500
Weighted Avg	0.91	0.91	0.91	500



Decision Tree(Hyperparameter Tuning)

Classification report for Decision Tree (Test set)

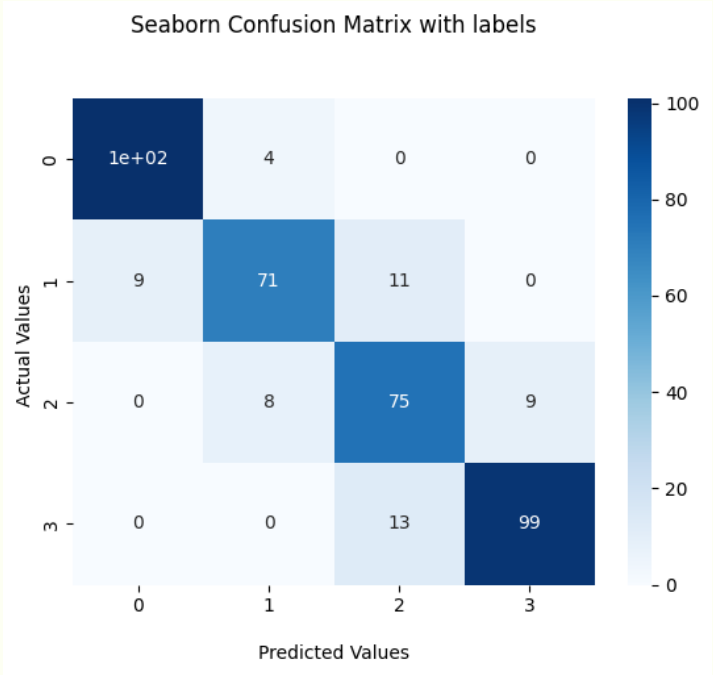
	Precision	Recall	F1-Score	Support
0	0.95	0.87	0.91	132
1	0.75	0.86	0.80	118
2	0.78	0.72	0.75	120
3	0.88	0.89	0.89	130
Accuracy			0.84	500
Macro Avg	0.84	0.84	0.84	500
Weighted Avg	0.84	0.84	0.84	500



Random Forest(Hyperparameter Tuning)

Classification report for Random Forest (Test set)

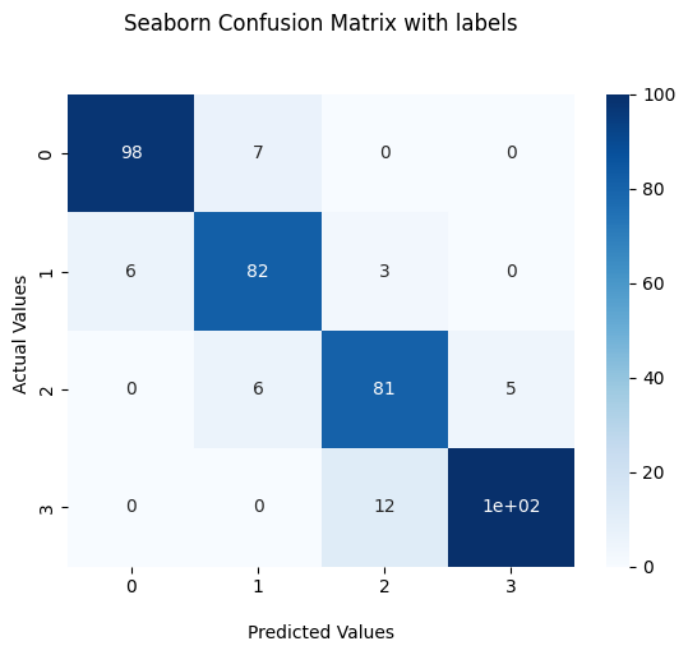
	Precision	Recall	F1-Score	Support
0	0.92	0.96	0.94	105
1	0.86	0.78	0.82	91
2	0.76	0.82	0.79	92
3	0.92	0.88	0.90	112
Accuracy			0.86	400
Macro Avg	0.86	0.86	0.86	400
Weighted Avg	0.87	0.86	0.86	400



xgBoost(Hyperparameter Tuning)

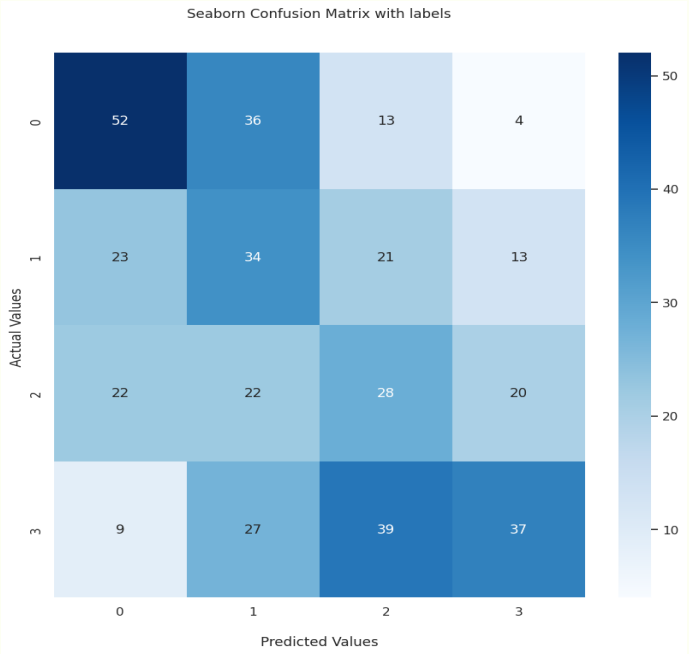
Classification report for xgBoost (Test set)

	Precision	Recall	F1-Score	Support
0	0.94	0.93	0.94	105
1	0.86	0.90	0.88	91
2	0.84	0.88	0.86	92
3	0.95	0.89	0.92	112
Accuracy			0.90	400
Macro Avg	0.90	0.90	0.90	400
Weighted Avg	0.90	0.90	0.90	400



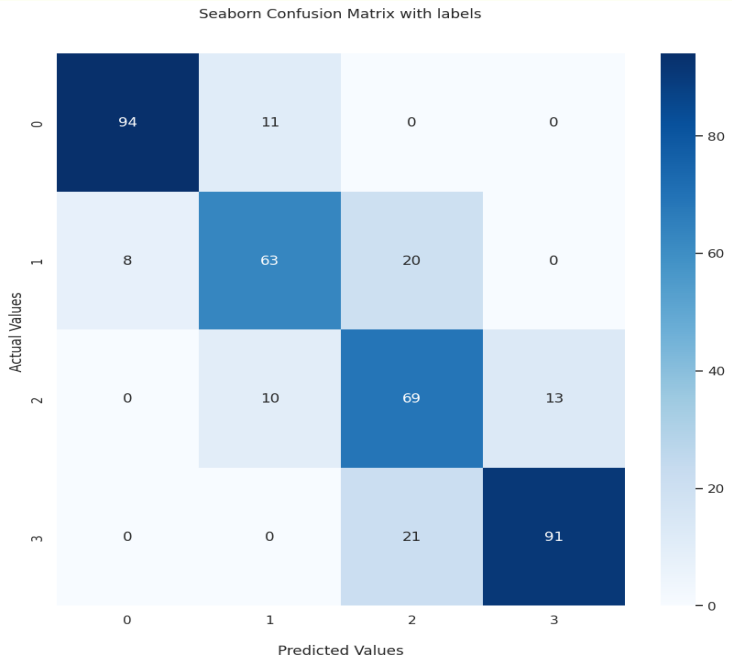
Knn classifier

	Precision	Recall	F1-Score	Support
0	0.92	0.90	0.91	105
1	0.75	0.69	0.72	91
2	0.63	0.75	0.68	92
3	0.88	0.81	0.84	112
Accuracy			0.79	400
Macro Avg	0.79	0.79	0.79	400
Weighted Avg	0.80	0.79	0.80	400



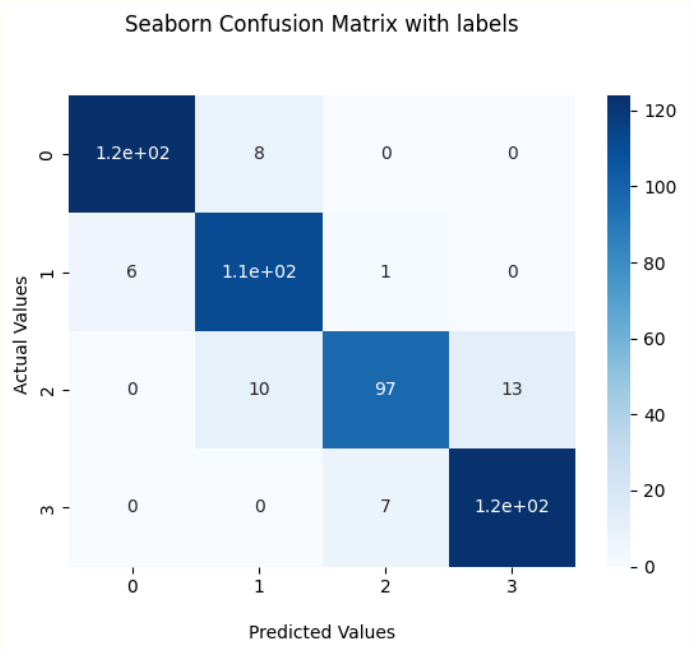
Naive Bayes

	Precision	Recall	F1-Score	Support
0	0.92	0.90	0.91	105
1	0.75	0.69	0.72	91
2	0.63	0.75	0.68	92
3	0.88	0.81	0.84	112
Accuracy			0.79	400
Macro Avg	0.79	0.79	0.79	400
Weighted Avg	0.80	0.79	0.80	400

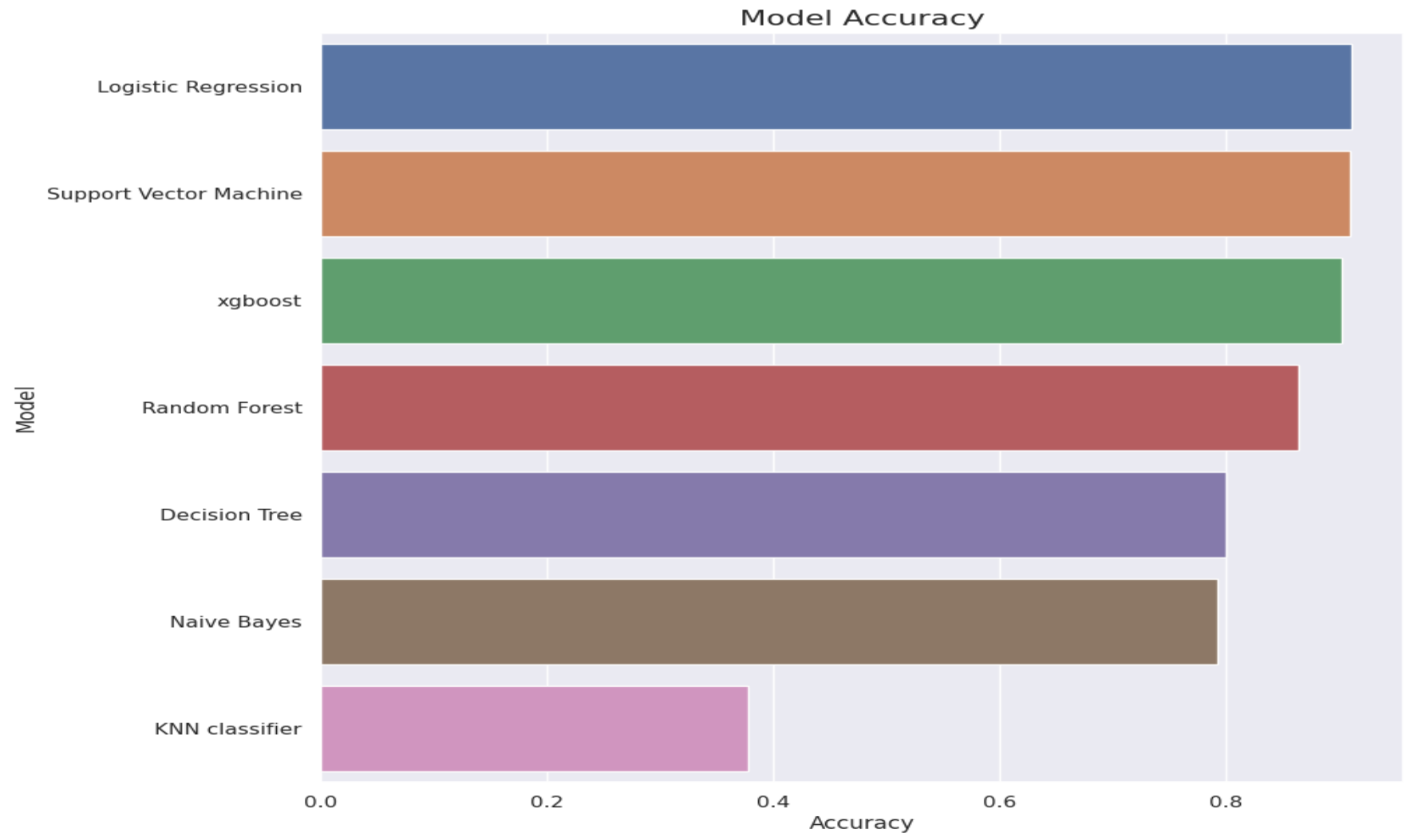


Support Vector Machine

	Precision	Recall	F1-Score	Support
0	0.95	0.94	0.95	132
1	0.86	0.94	0.90	118
2	0.92	0.81	0.86	120
3	0.90	0.95	0.92	130
Accuracy			0.91	500
Macro Avg	0.91	0.91	0.91	500
Weighted Avg	0.91	0.91	0.91	500

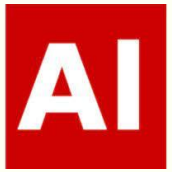


Model Performance



Conclusions of Modelling

- The given dataset was very clean it has no null values, no duplicates, and no outliers
- Half the devices have Bluetooth, and half don’t there is a gradual increase in battery as the price range increases Ram has continuous increase with price range while moving from Low cost to Very high cost.
- Costly phones are lighter.
- RAM, battery power, pixels played more significant role in deciding the price range of mobile phone.
- Form all the above experiments we can conclude that Logistic Regression>Support Vector Machine>XGboost>Random Forest>Decision Tree> Naive Bayes>KNN with using hyperparameters will got the best results.



Thank You