Capstone Project

Mobile Price Range Prediction

Supervised Machine Learning (Classification)

INDEX

- Problem Statement
- Data Description
- Data Wrangling
- EDA (Exploratory Data Analysis)
- Feature Engineering
- Model Selection and Evaluation
- Evaluation of models
- Feature Importance
- Challenges
- Conclusion.

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.
- 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 (e.g.:- 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.
- The main objective of this project is to build a model which will classify the price range of mobile phones based on the specifications of mobile phones.

❖ Data Description

AI

Total Rows= 2000 , Total features=21

- **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.**

❖ Data Description

- Pc - Primary Camera mega pixels.
- Px_height and Px_width - Pixel Resolution Height and width.
- Ram - Random Access Memory in Mega Bytes.
- Sc_h and Sc_w - Screen Height and 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.

Target variable:

- Price_range - This is the target variable with value of 0(low cost),1(medium cost),2(high cost) and3(very high cost).

❖ Data Wrangling

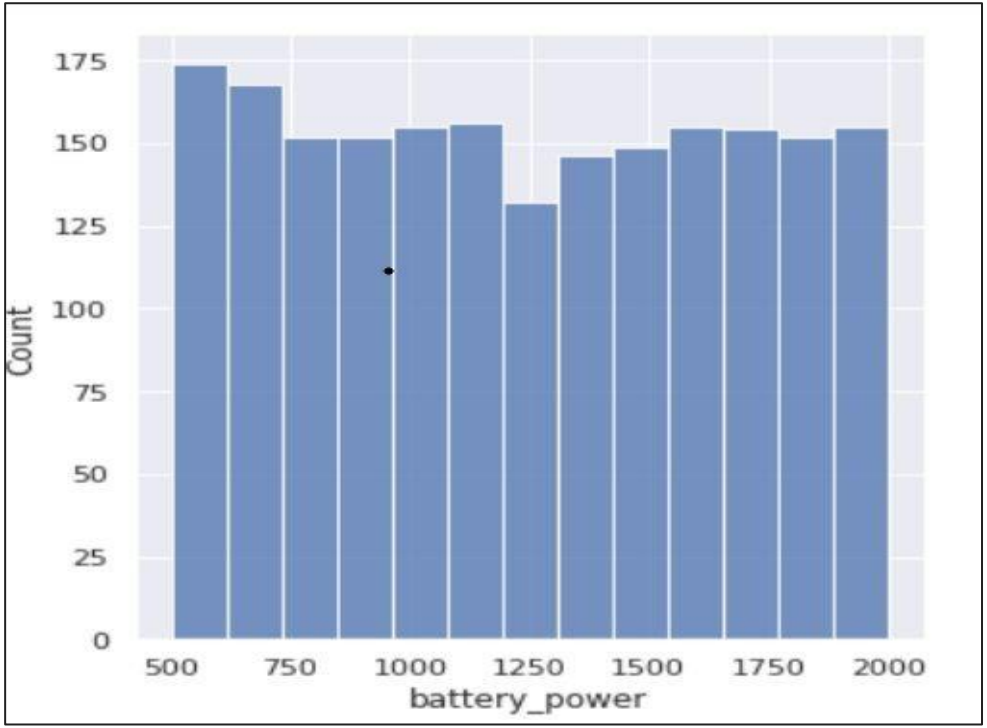
➢ Handling missing values present in data

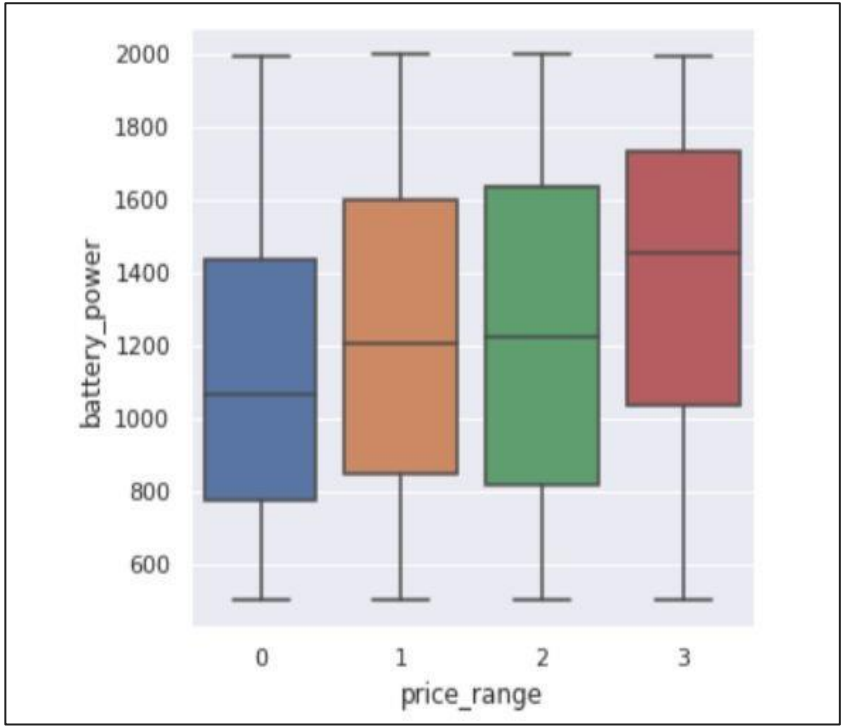
	count	mean	std	min	25%	50%	75%	max
sc_width	2000.0	5.76700	4.356398	0.0	2.00	5.0	9.00	18.0
px_height	2000.0	645.10800	443.780811	0.0	282.75	564.0	947.25	1960.0

- ➢ From above information, min screen width & min pixel height is zero. So that means this information contains missing values so let's check there respective counts.
- ➢ Total phones with sc_w = 0 is 180
- ➢ Total phones with px_height = 0 is 2
- ➢ Where there is sc_W and px_height is zero ,assigning mean value to that value in order to handle missing values.
- ➢ In This data set there is not any duplicate values present.

Exploratory Data Analysis

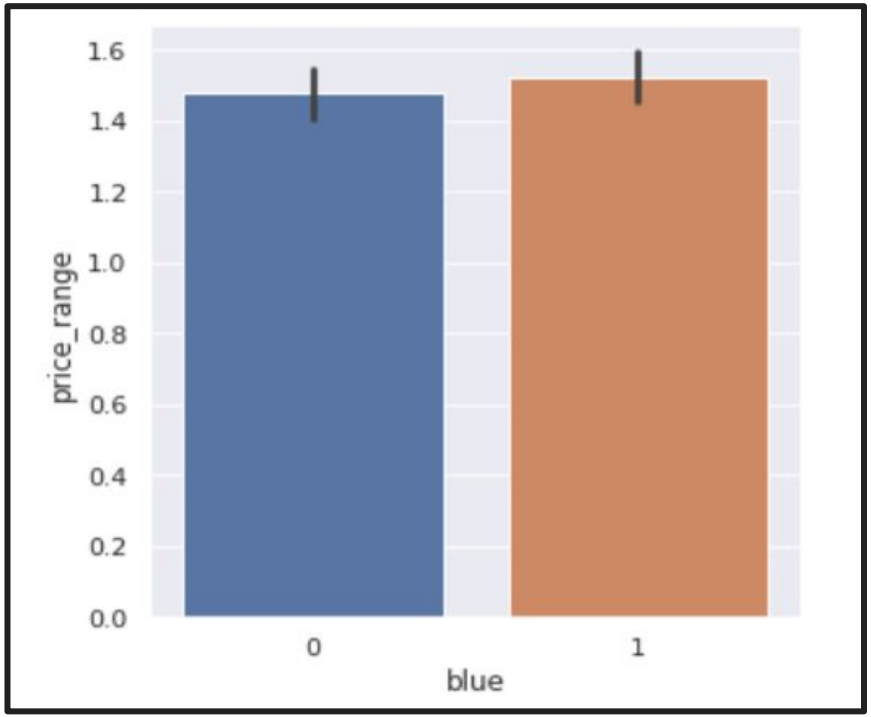
This Graph shows that as battery power increase there is gradual increase in mobile price range.



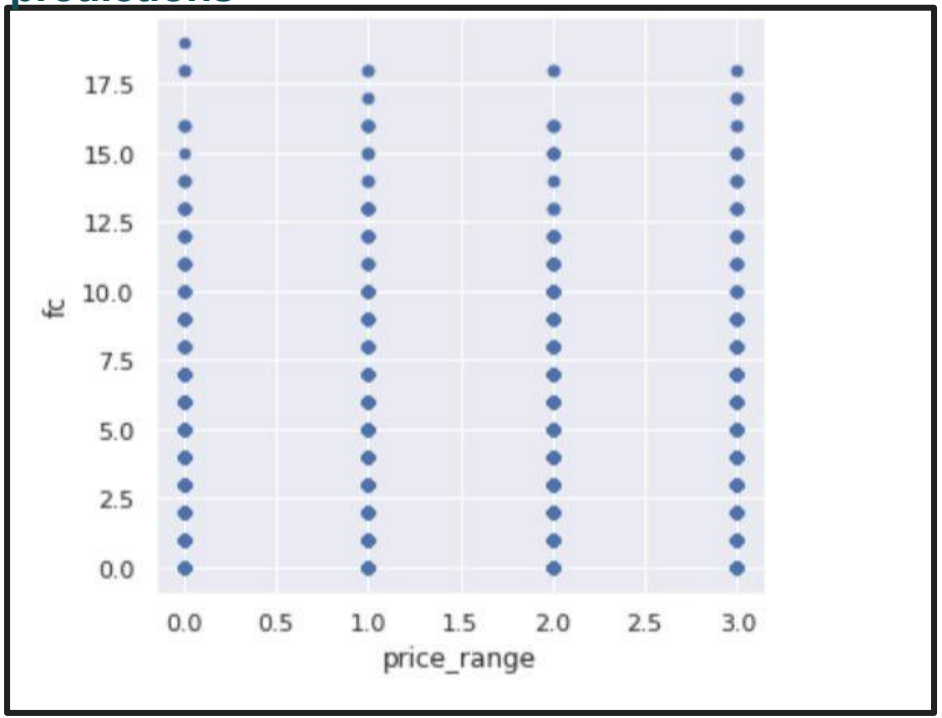


Exploratory Data Analysis

Having bluetooth or not do not affect the price of mobile

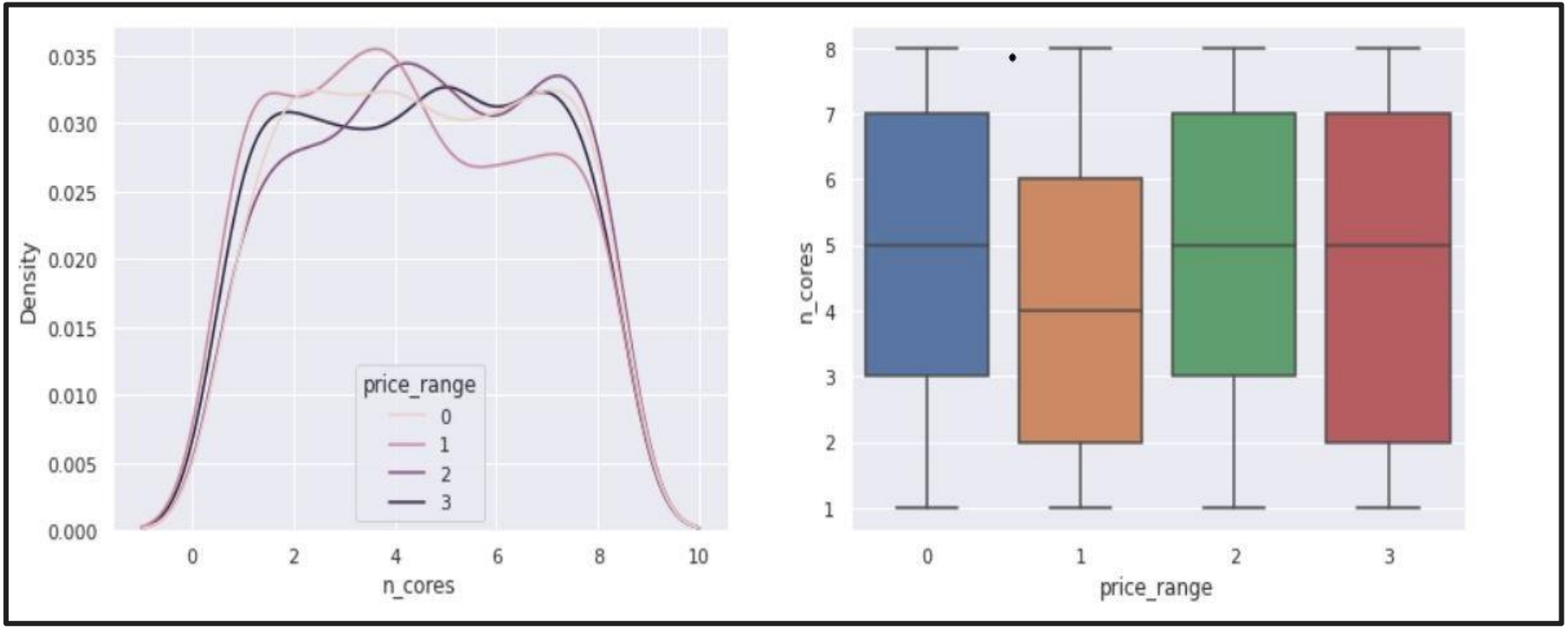


Front camera megapixel distribution is almost similar along all the price ranges variable, it may not be helpful in making predictions



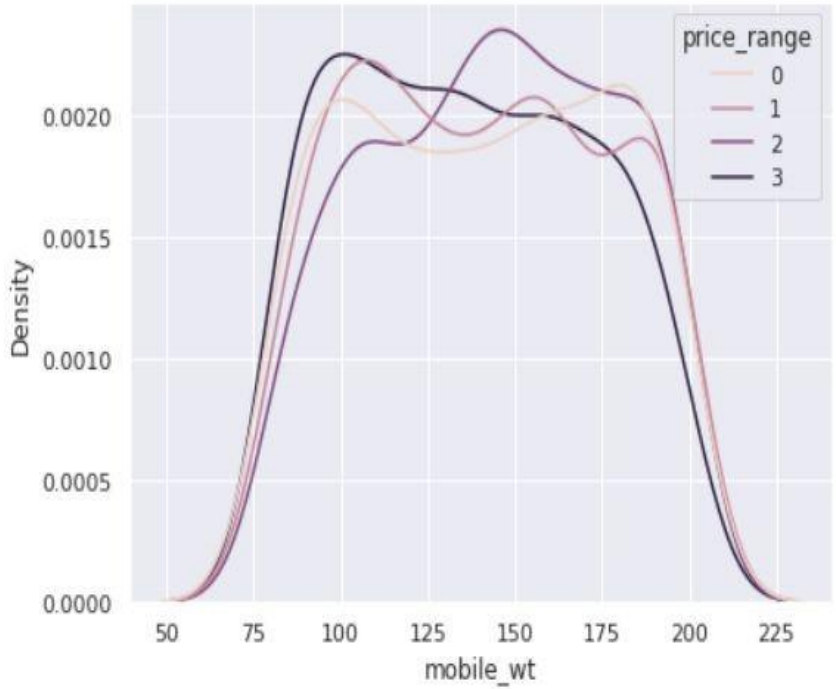
Exploratory Data Analysis

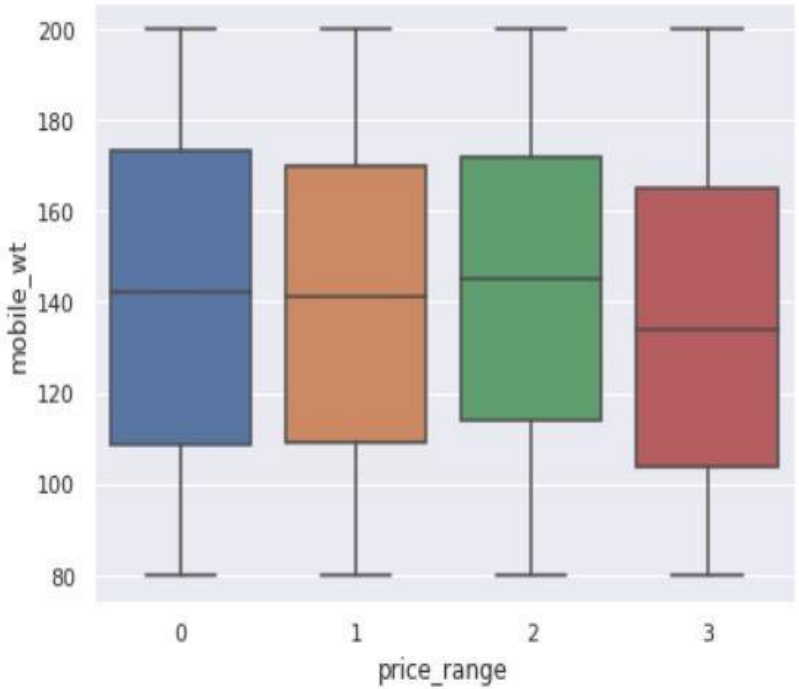
No of core Processors are showing a little variation along the target categories that helps in price prediction.



Exploratory Data Analysis

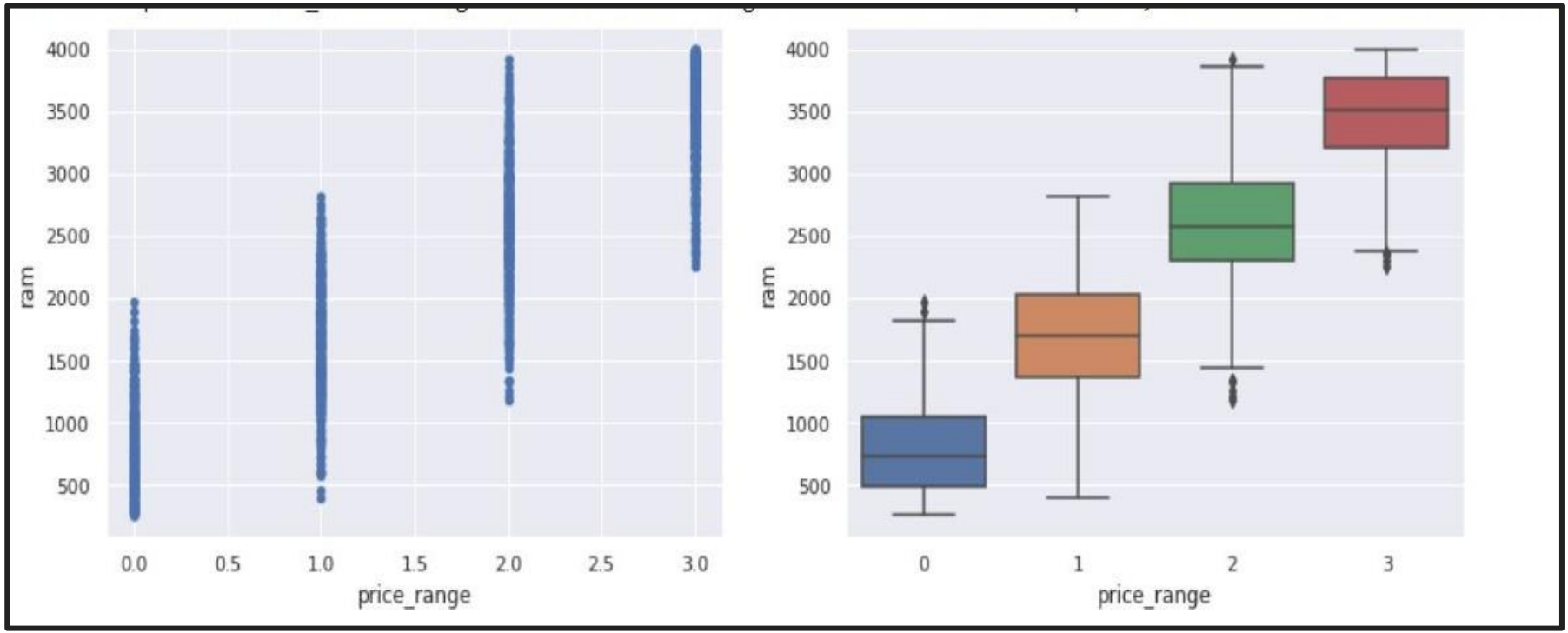
Costly phones are lighter.mobile weight affect the pricing of mobile





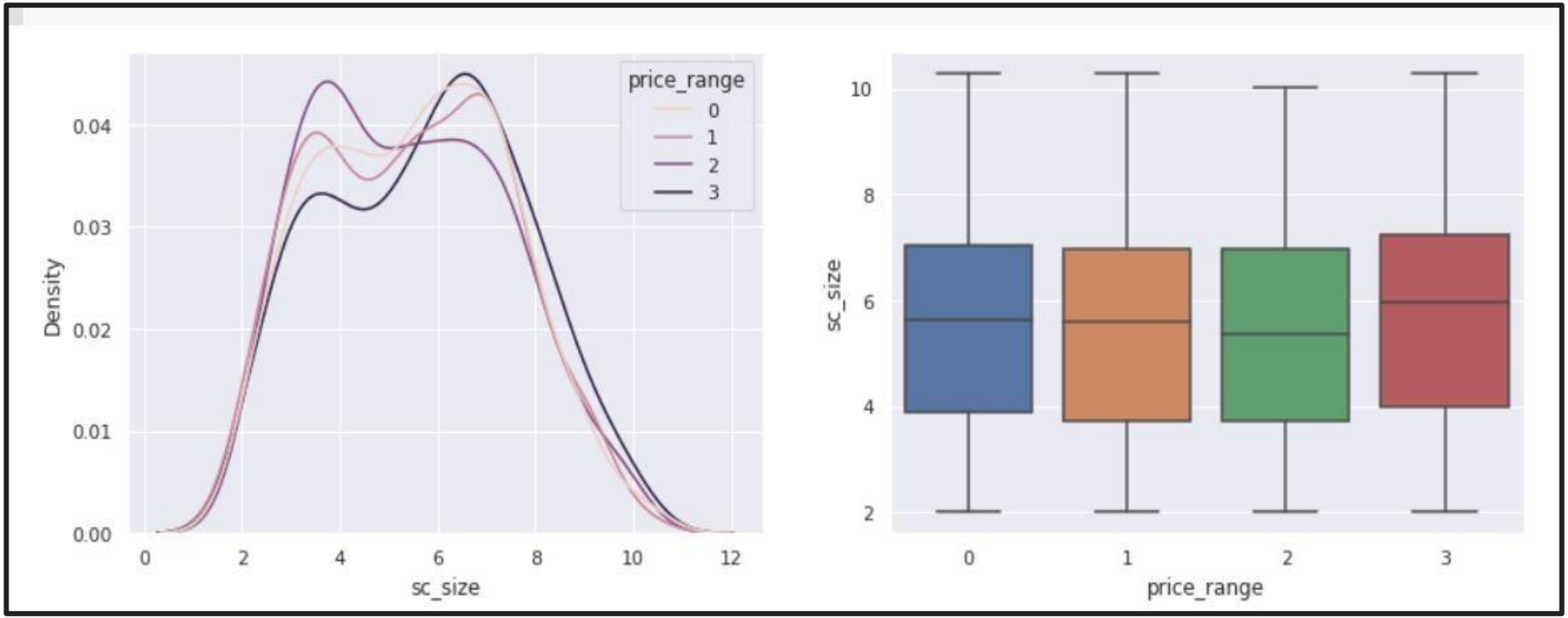
Exploratory Data Analysis

- 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.



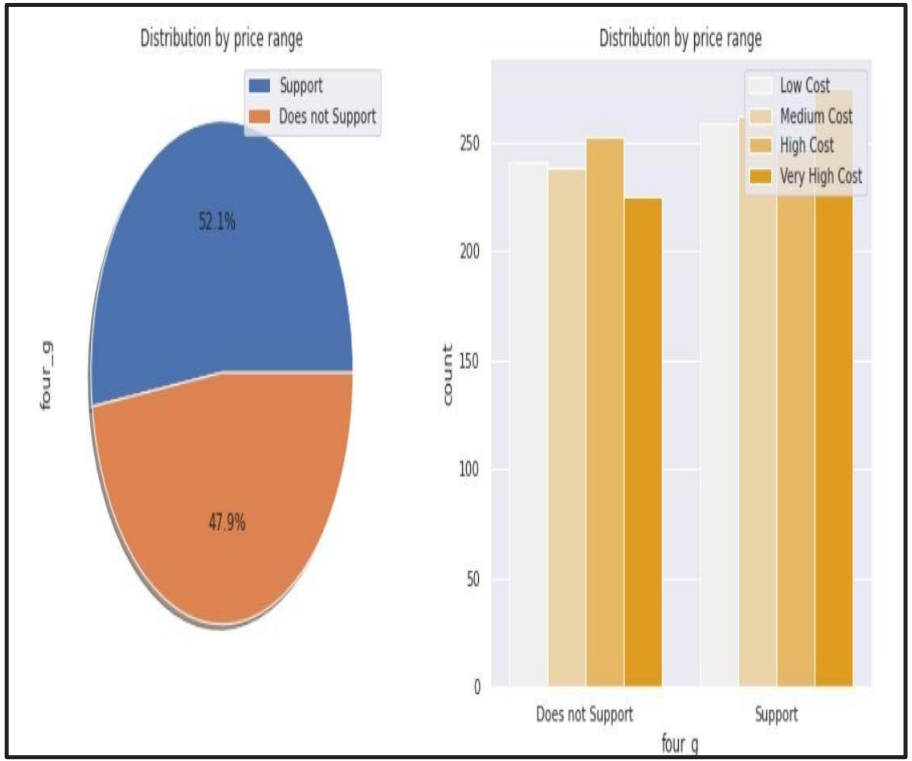
Exploratory Data Analysis

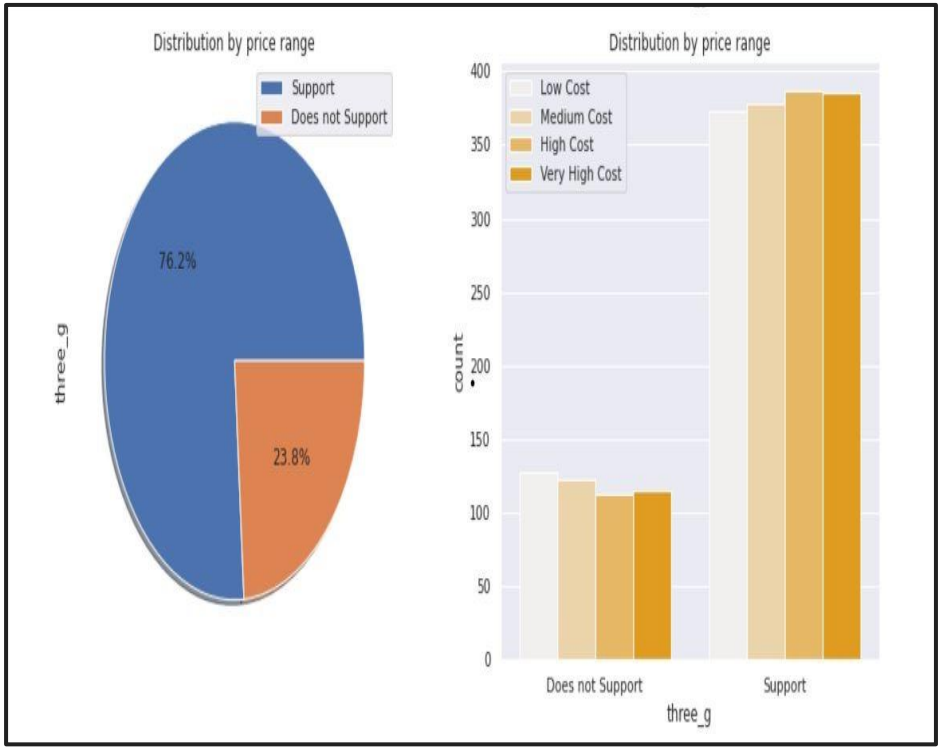
Screen Size shows little variation along the target variables. This can be helpful in predicting the target variable that is mobile price.



Exploratory Data Analysis

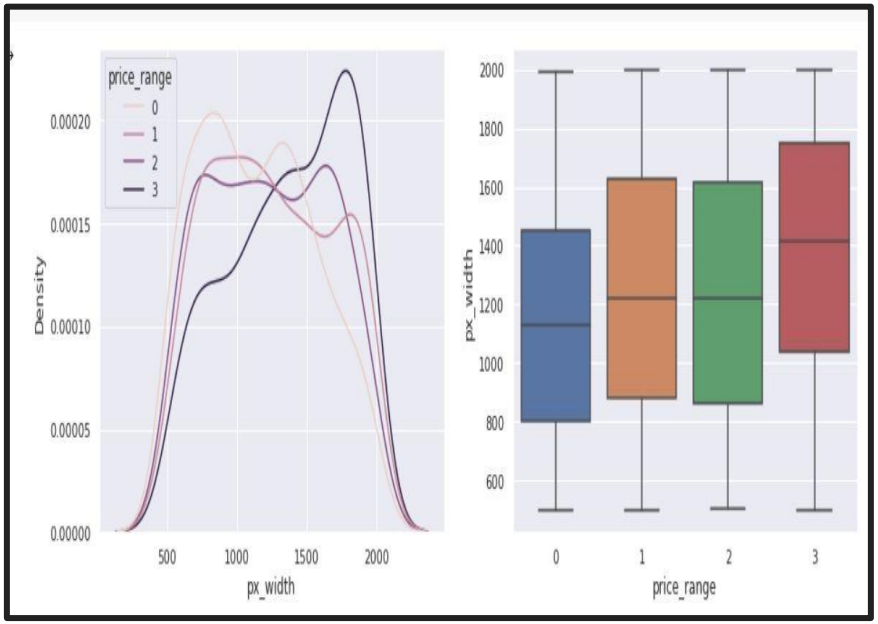
- 50% of the phones support 4_g and 76% of phones support 3_g
- feature 'three_g' play an important feature in prediction of mobile price.

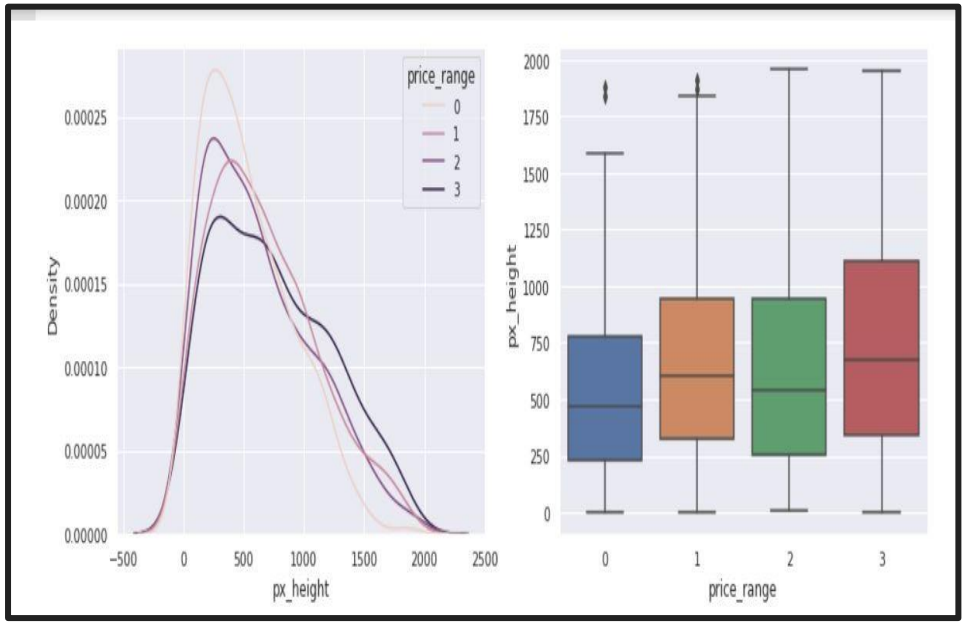




Exploratory Data Analysis

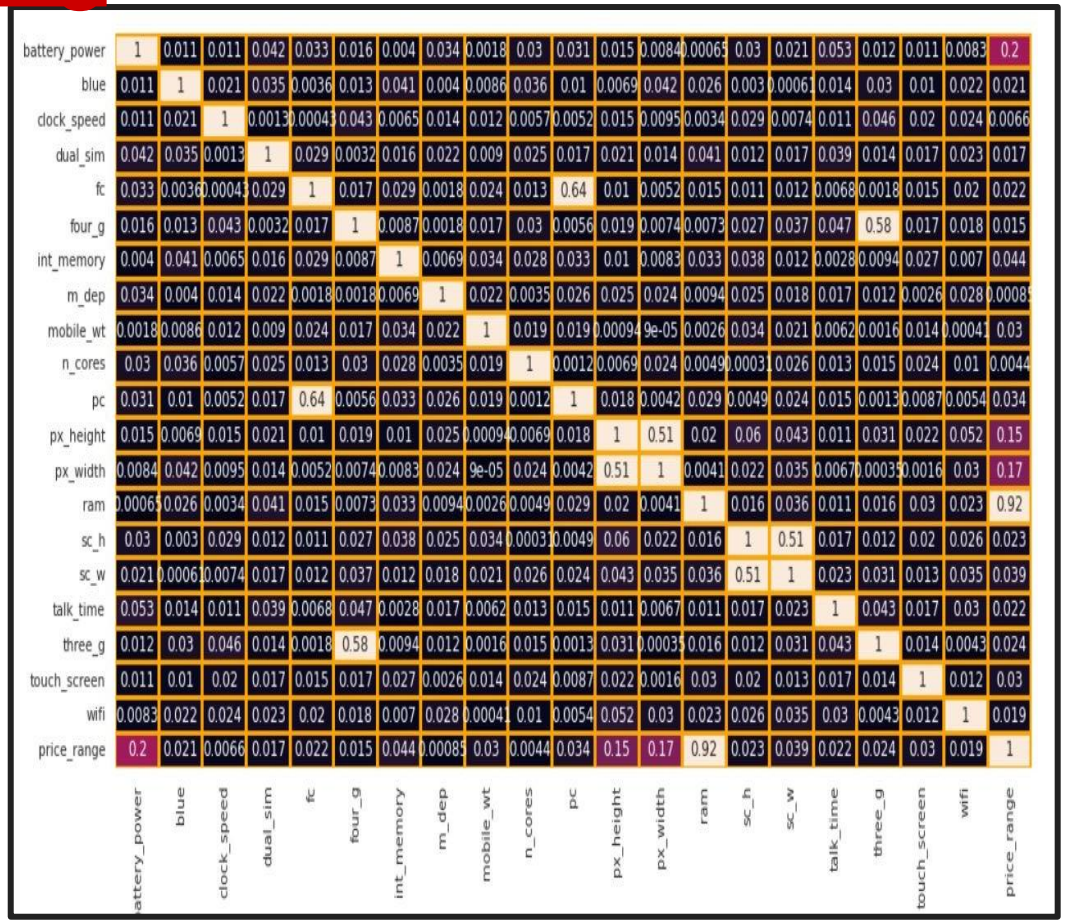
- Pixel height and pixel width shows little variation in mobile price range.
- There is not a continuous increase in pixel width & height as we move from Low cost to Very high cost. Mobiles with 'Medium cost' and 'High cost' has almost equal pixel width & height





Feature Engineering

- 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 features ('pc', 'fc') and ('px_width', 'px_height') and (sc_w,sc_h).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.
- Also, if sc_h increases, sc w also increases, that means the overall screen size increase. We can replace these two features with one feature i.e screen_size



❖ Feature Engineering

- Similarly the screen size of the phone is expressed in Inches.
- We have columns **'sc_h'** and **'sc_w'** out of which we have created a new feature **'Screen_size'** which is diagonal length of the screen.

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:

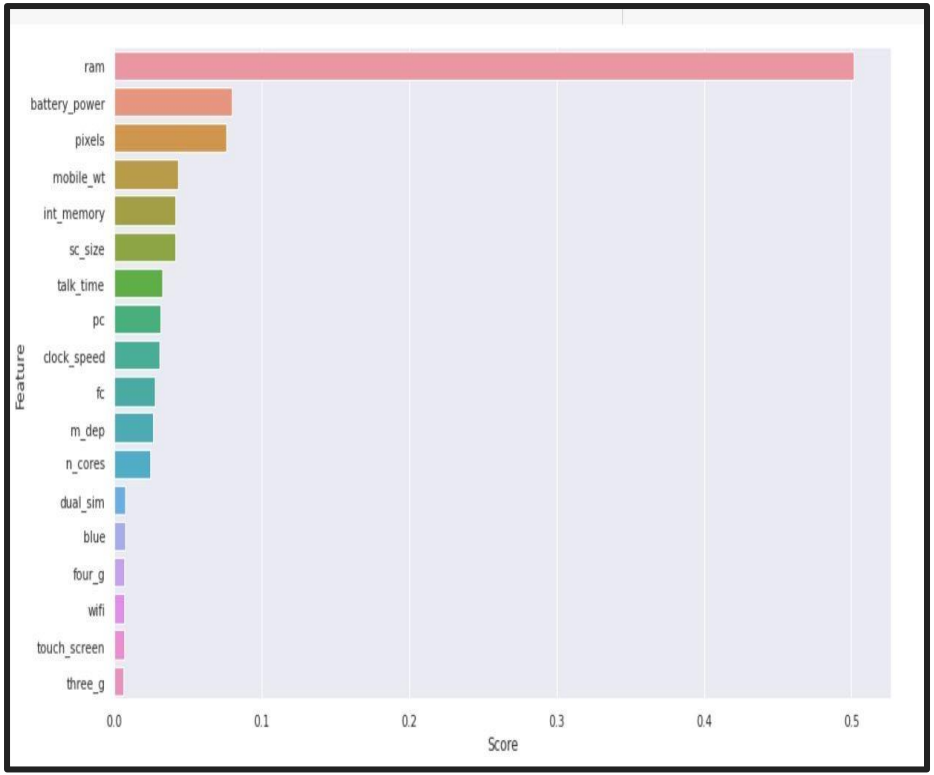
Algorithms	Training Set		Test set	
	Accracy score (%)	Recall score(%)	Accracy score (%)	Recall score(%)
Logistic Regression	92%	92%	89%	89%
Random forest	100%	100%	87%	87%
Random forest (hyper tunning)	95%	95%	84%	84%
K Nearst Neighbour	100%	100%	44%	44%
XGBoost	98%	98%	89%	89%
Support Vector Machine	94%	94%	90%	90%
Gradient Boosting	99%	99%	90%	90%
Decision Tree	88%	88%	84%	84%

- Best model came out to be Support VM.
- XG boost can be considered as the second most good model.
- KNN performed very worst.

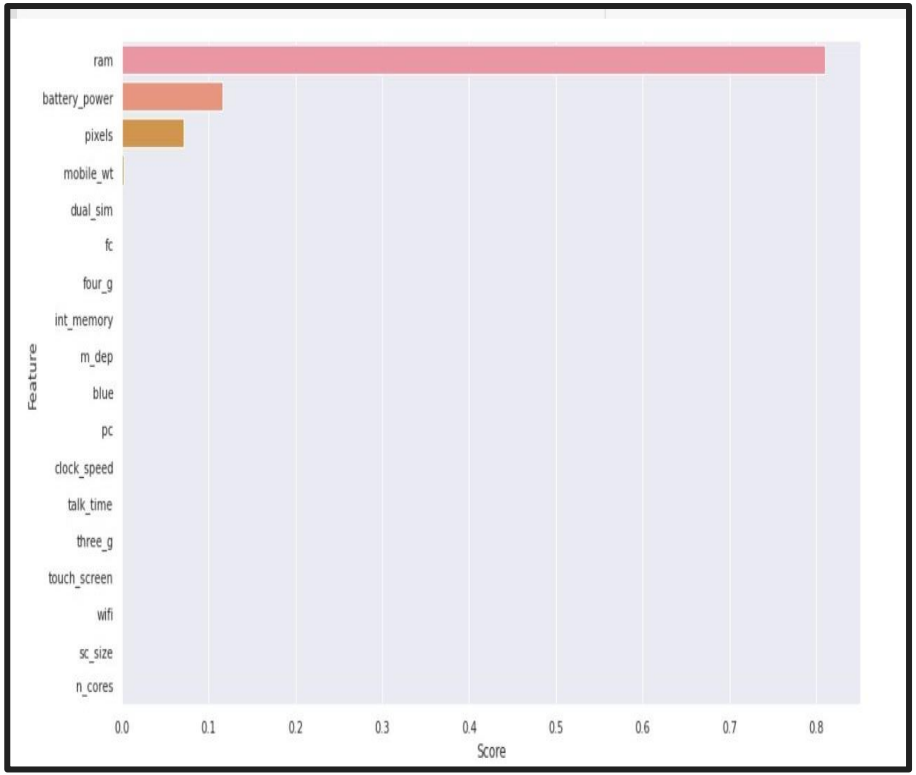
❖ Evaluation of models:

❖ Feature Importance

Random forest

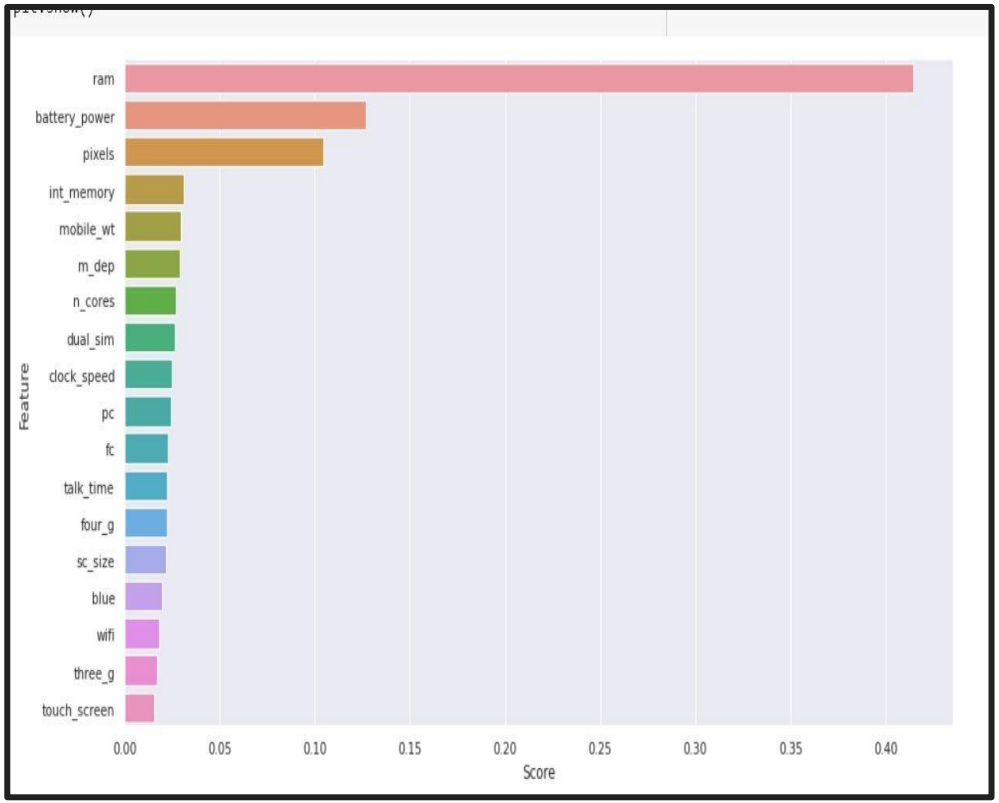


Decision Tree

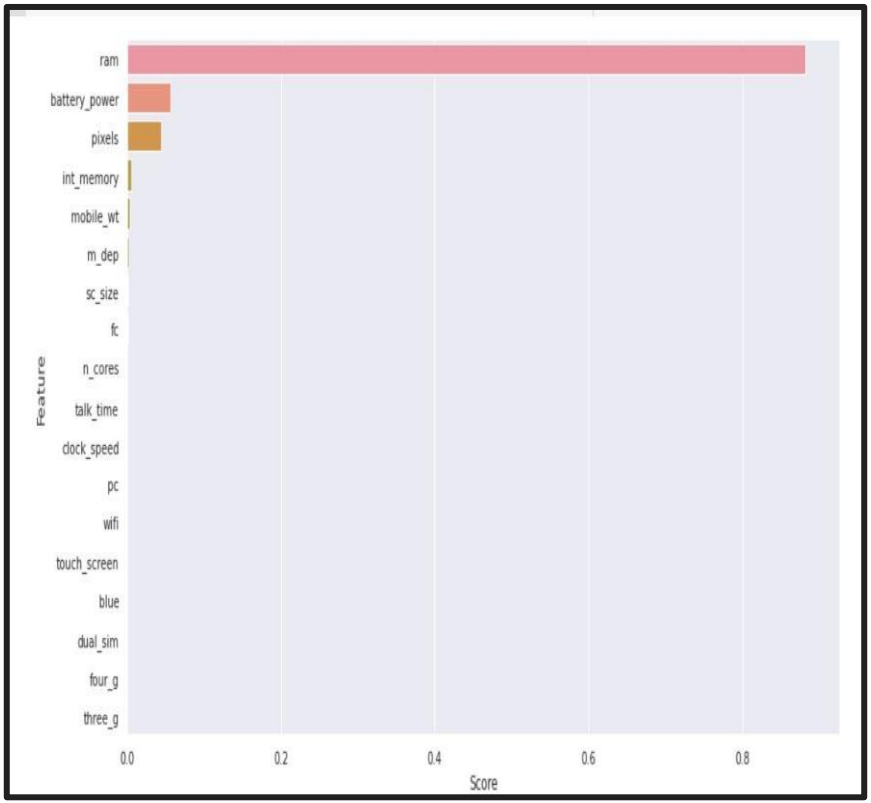


❖ Feature Importance

XGBoost



Gradient Boosting



❖ Challenges

- We performed “Hypothesis driven EDA” based on domain, but unlikely most of our hypothesis got rejected by our data.
- Most of the models are not able to get good accuracy for each class of target variable.
- We hit a ceiling at 94% accuracy using a single model.

Conclusions:

- 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 with 94% train accuracy and 90 % test accuracy.
- XG boost is the second best good model which gave good performance 98% train accuracy and 89% 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,Screen size,mobile weight contributed the most while predicting the price range.

THANK YOU