

# Capstone Project-3

## Mobile Price Range Prediction

(Supervised Machine Learning (Classification))

BY

**Prasad Kanagi**

# 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.

# Data Description: 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.

# Data Description:

- ❖ **Mobile\_wt** - Weight of mobile phone.
- ❖ **N\_cores** - Number of cores of processor.
- ❖ **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.

# Data Wrangling:

## Handling Mismatch values in data.

### ▼ Some Basic Observations:

We can see that `sc_width` and `px_height` has minimum value 0, which is not possible in any mobile. We need to handle this mismatch.

```
[ ] # Checking How many observations having screen width value as 0.  
    print(data[data['sc_w']==0].shape[0])
```

180

```
▶ # Checking How many observations having px_height value as 0.  
   print(data[data['px_height']==0].shape[0])
```

2

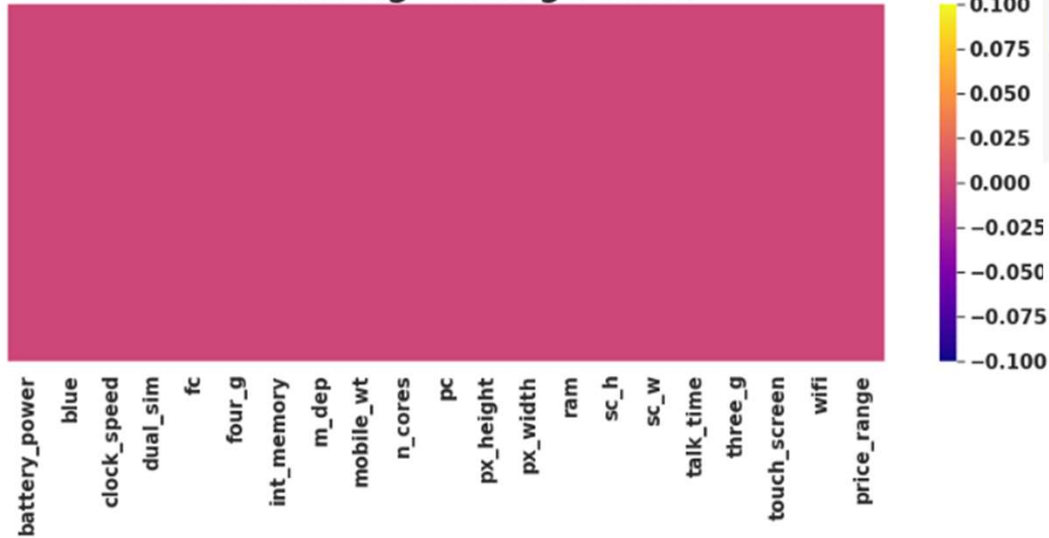
```
▶ # As there are only 2 observations having px_height=0. so we will drop  
   data=data[data['px_height']!=0]
```

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

# Data Wrangling:

AI

Visualising Missing Values



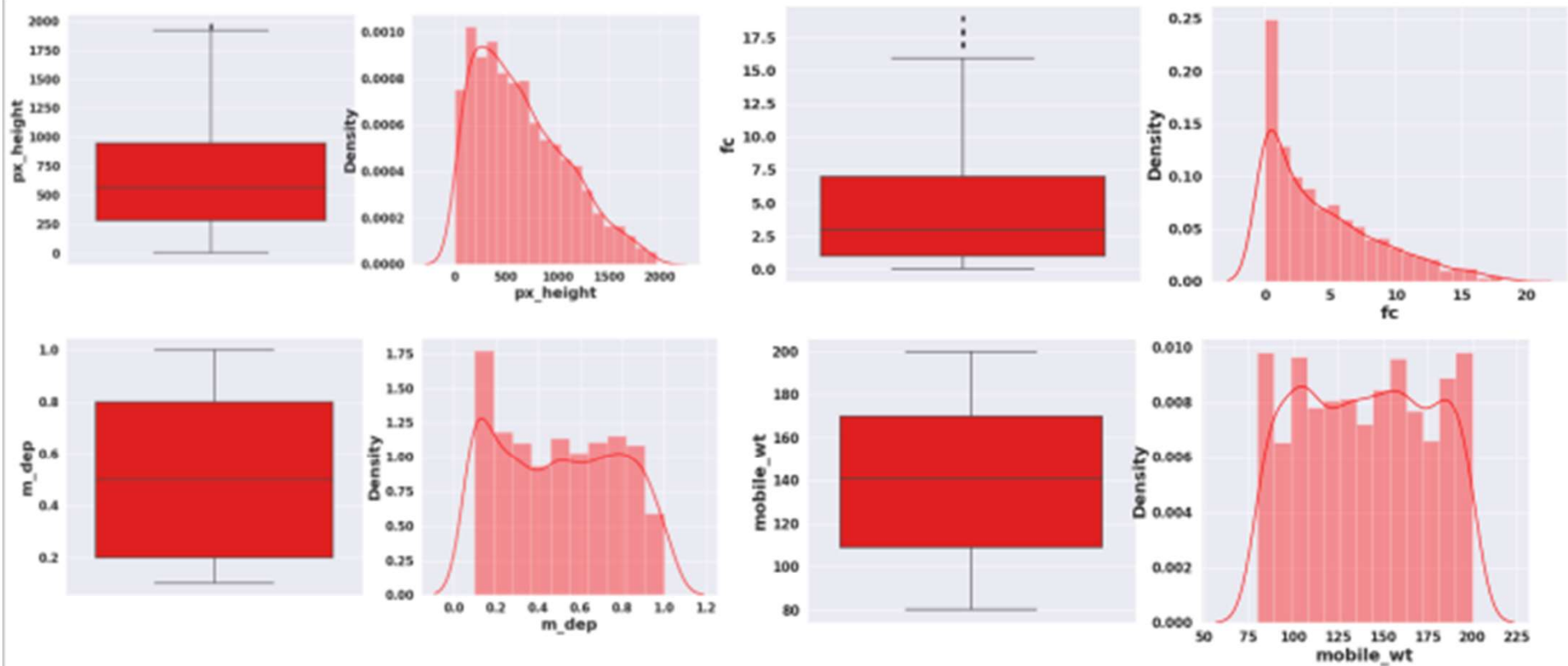
```
[ ] # Checking Duplicate values in data set.  
print(f' We have {data.duplicated().sum()} duplicate values in dataset.')
```

We have 0 duplicate values in dataset.

- ❖ Zero Missing values after handling mismatch from the data.
- ❖ 0 duplicates.

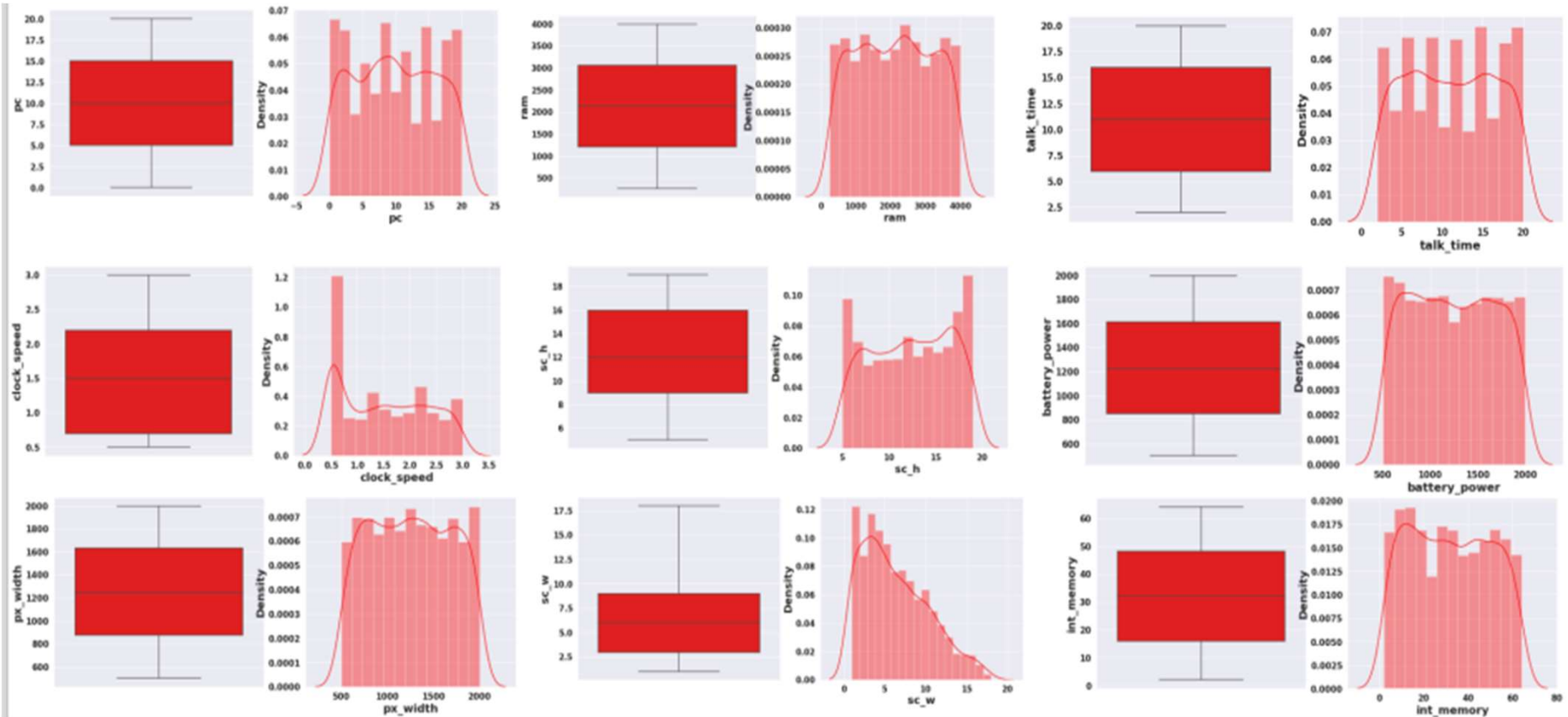
# Data Wrangling:

Checking outliers and Distribution of numerical variables



# Data Wrangling:

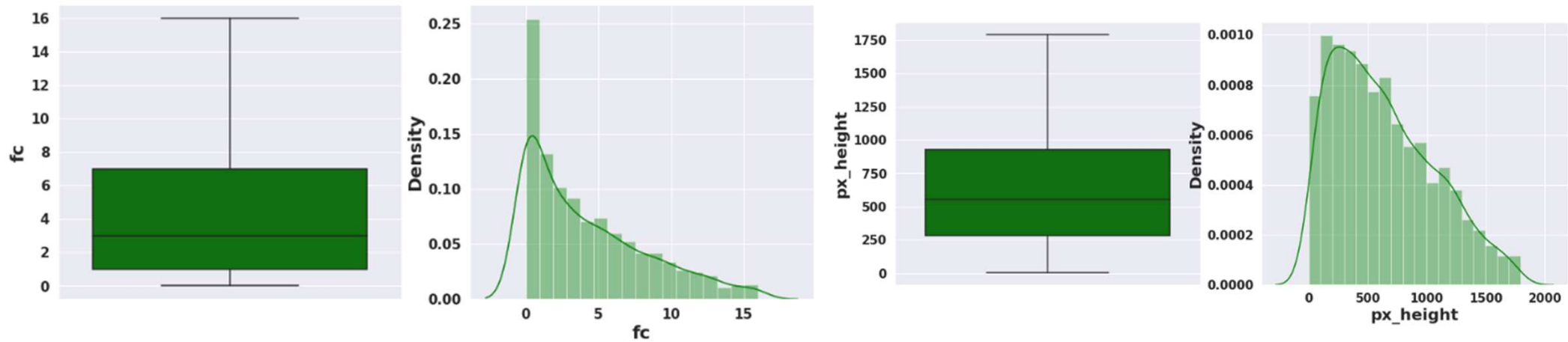
AI





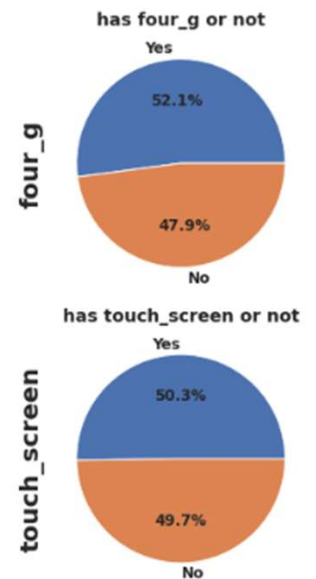
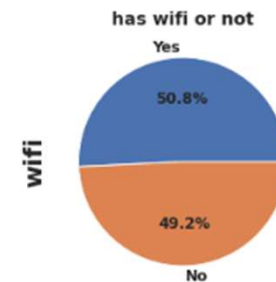
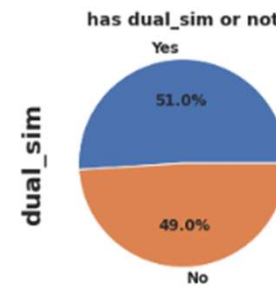
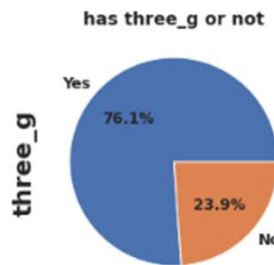
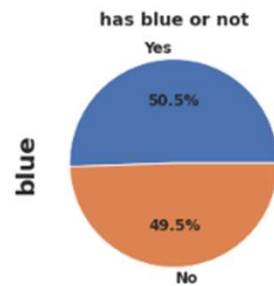
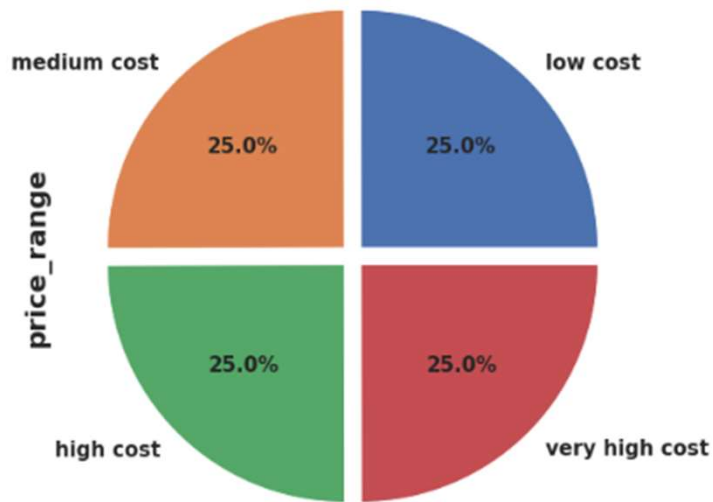
# Data Wrangling:

After removal of outliers:



# EDA (Exploratory Data Analysis):

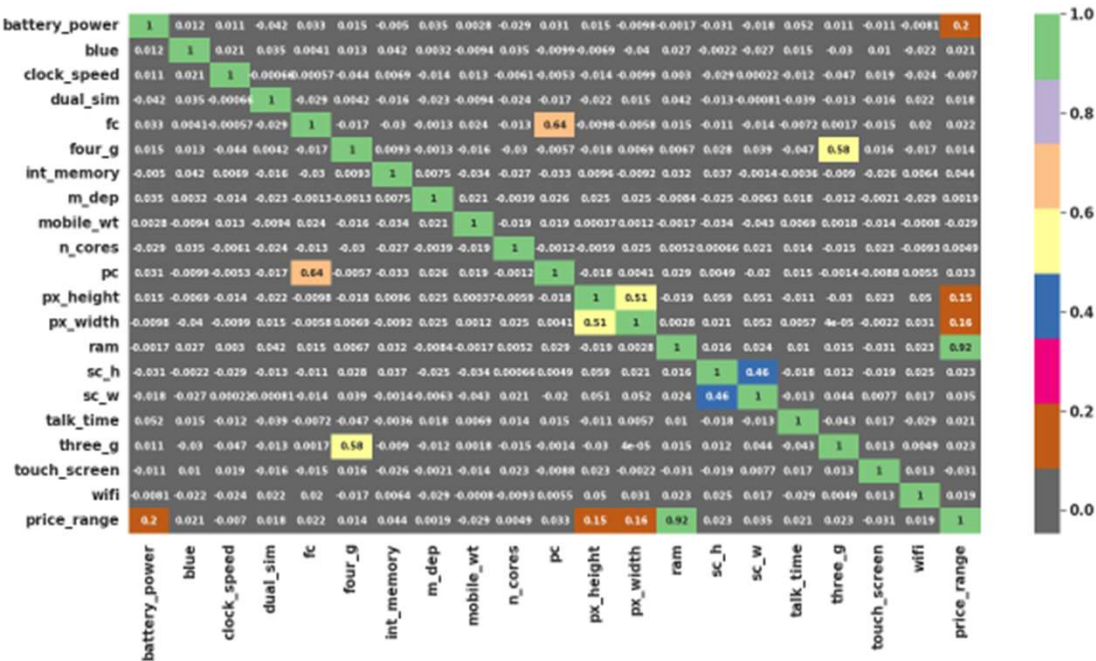
## Univariate Analysis:



- ❖ 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 .

# EDA (Exploratory Data Analysis):

## Bivariate and Multivariate Analysis: Correlation of independent variable with target variable

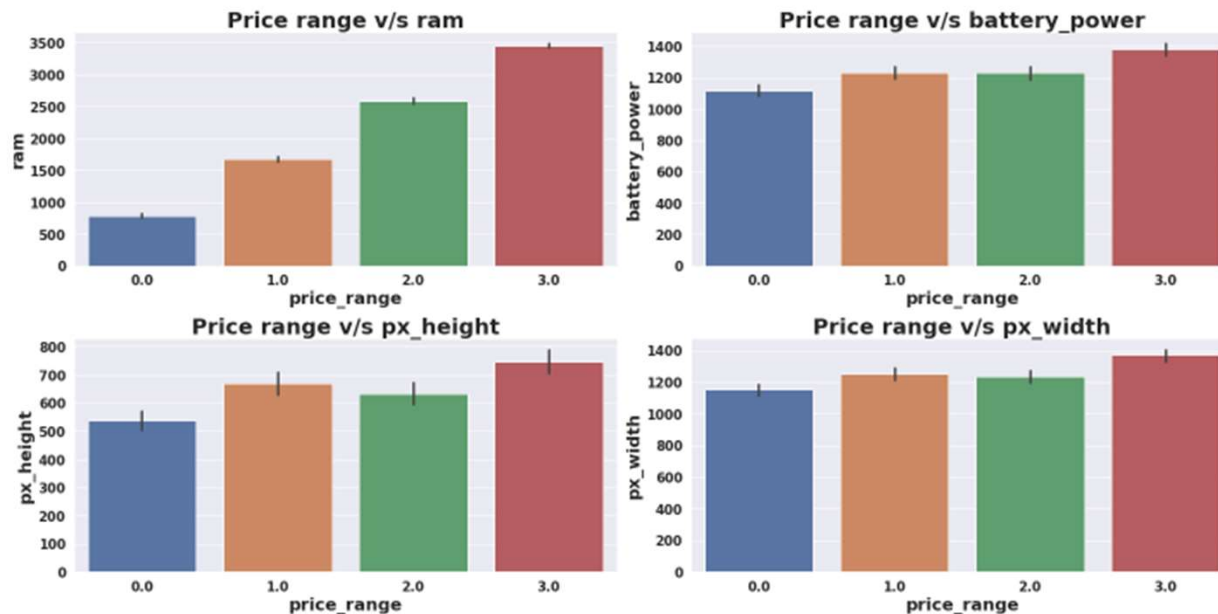


- ❖ primary camera correlated. i.e pc and front camera fc are positively correlated.
- ❖ sc\_h and sc\_w are positively correlated.

- ❖ RAM has strong positive correlation with the Price range and we know that Mobiles with high RAM are very costly. Thus RAM increases price range also increase.
- ❖ Battery power also has positive correlation with the price range. Generally mobiles having high prices comes with good battery power.
- ❖ Also px\_height and px\_width (Pixel Resolution Height and width) are positively correlated. Generally High price range mobiles have good resolutions.

# EDA (Exploratory Data Analysis):

## Bivariate and Multivariate Analysis:

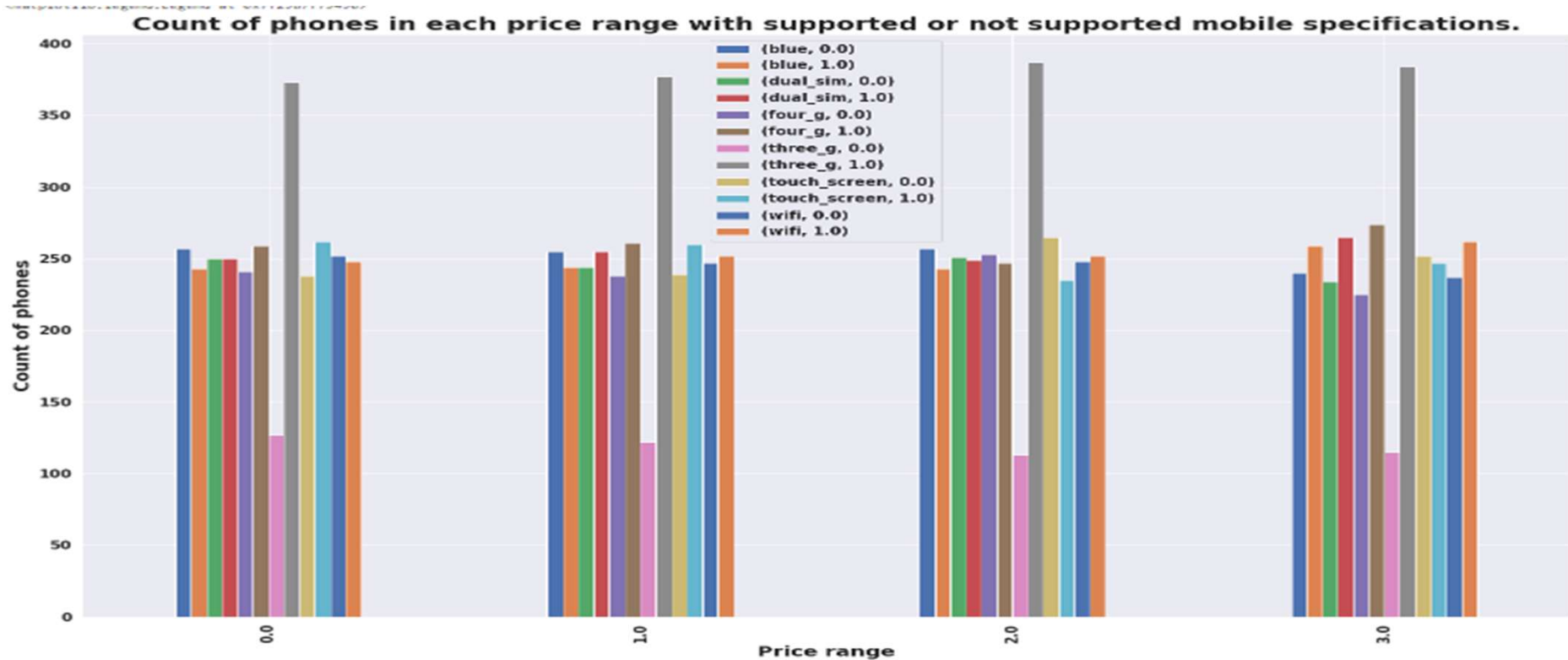


- ❖ Mobiles with battery power more than 1300 mAh has very high cost. And Mobiles with battery power between 1200 and 1300 mAh falls under medium and high cost category.

- ❖ Mobiles 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 more than 700 pixel height and width more than 1300 has very high cost.

# EDA (Exploratory Data Analysis):

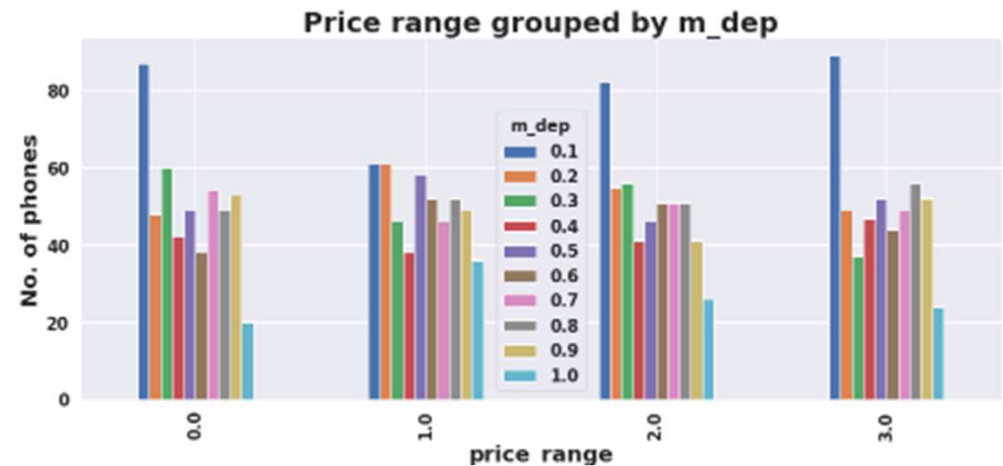
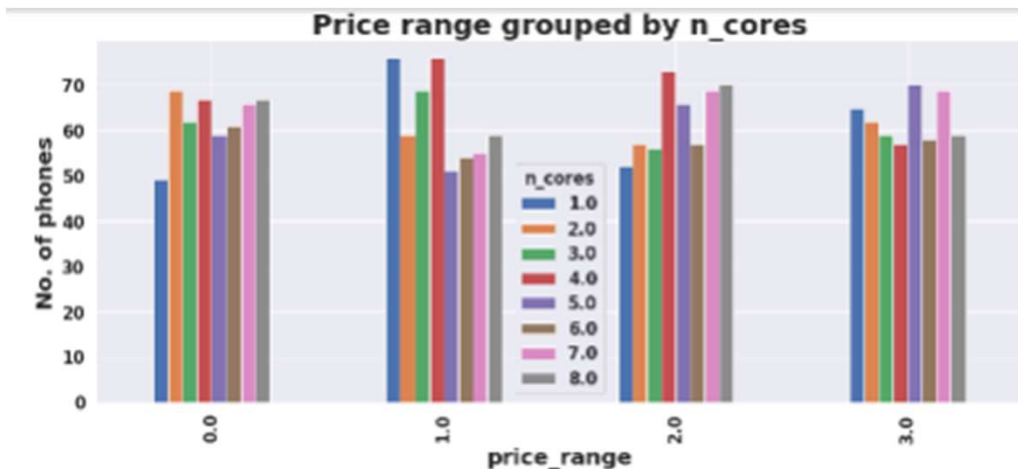
## Bivariate and Multivariate Analysis:



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

# EDA (Exploratory Data Analysis):

## Bivariate and Multivariate Analysis:

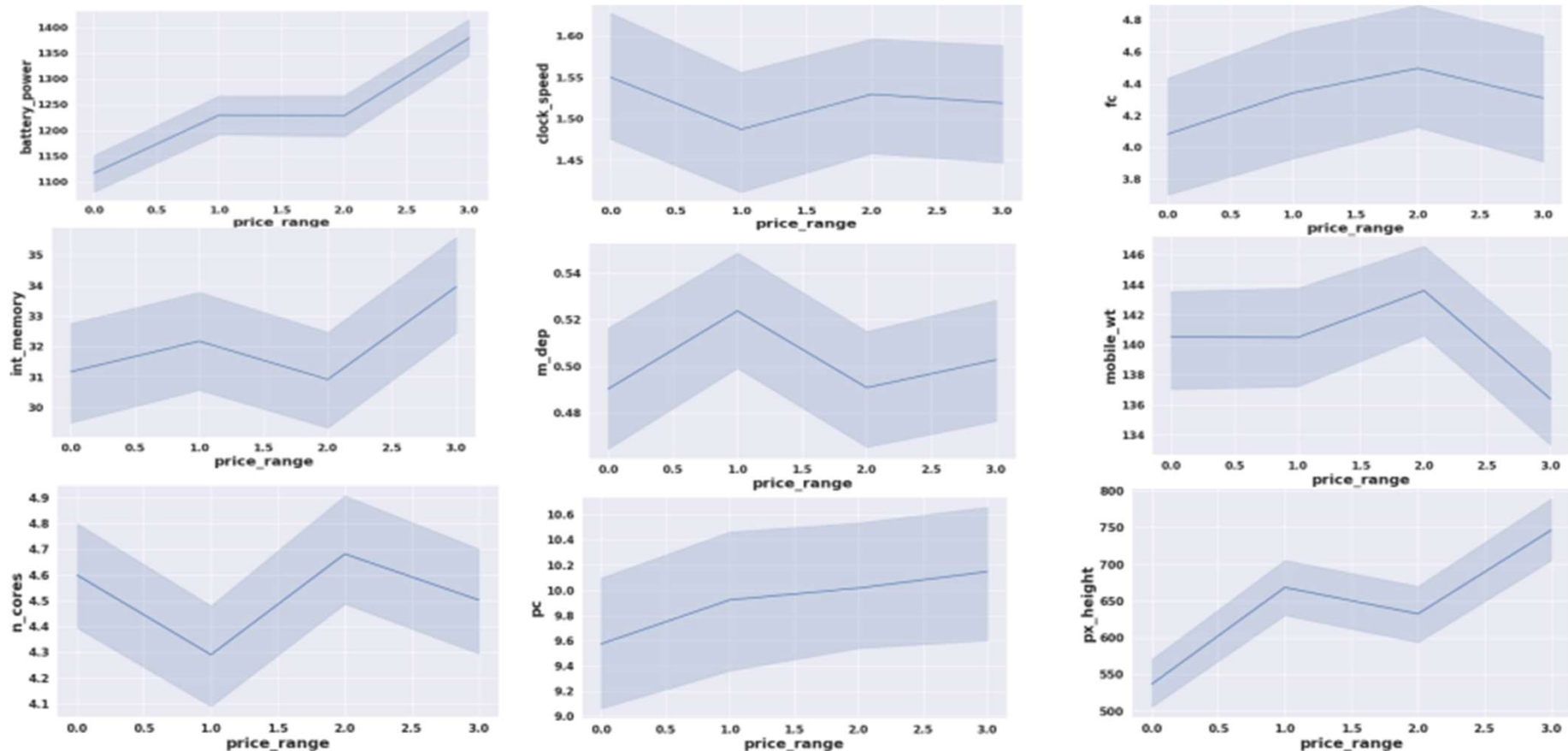


**There are very few mobiles in price range 0 and 1 with lesser no of cores.**

- ❖ 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.

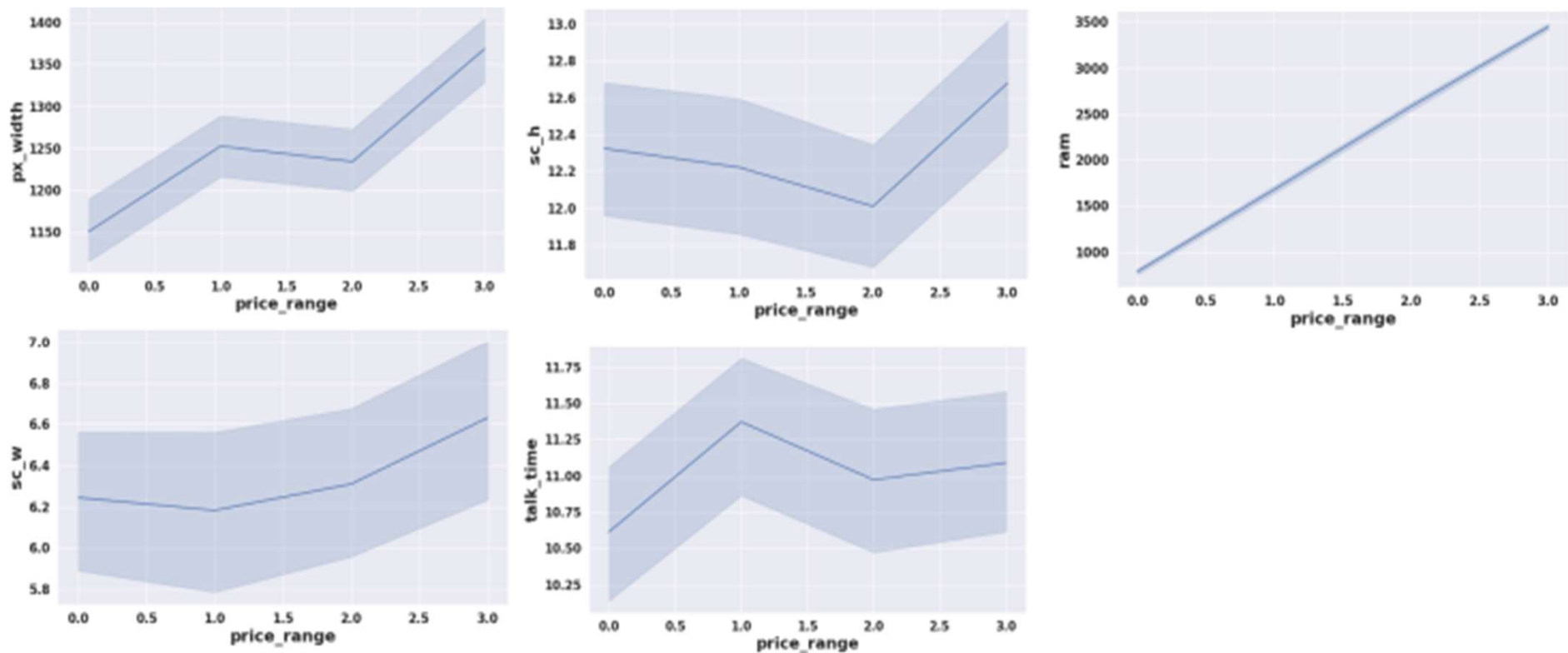
# EDA (Exploratory Data Analysis):

**Bivariate and Multivariate Analysis:** : Different trends of price range v/s other features



# EDA (Exploratory Data Analysis):

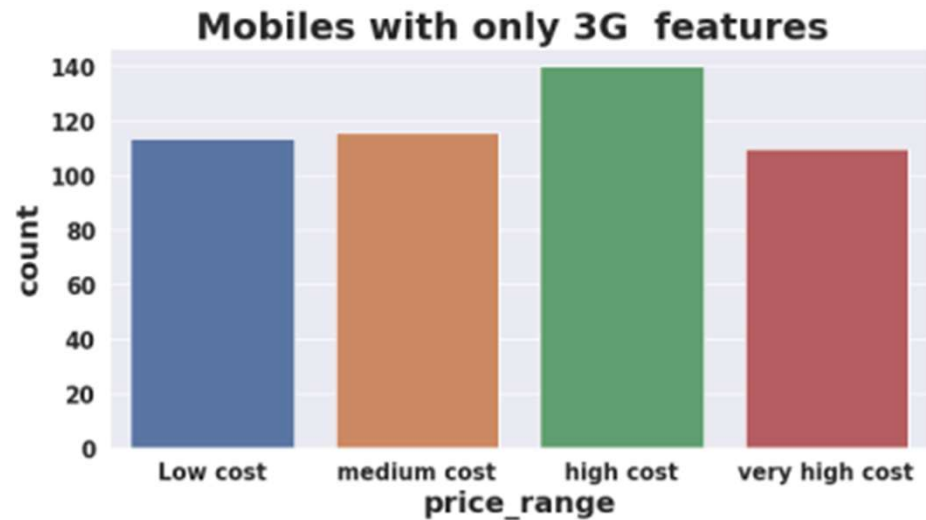
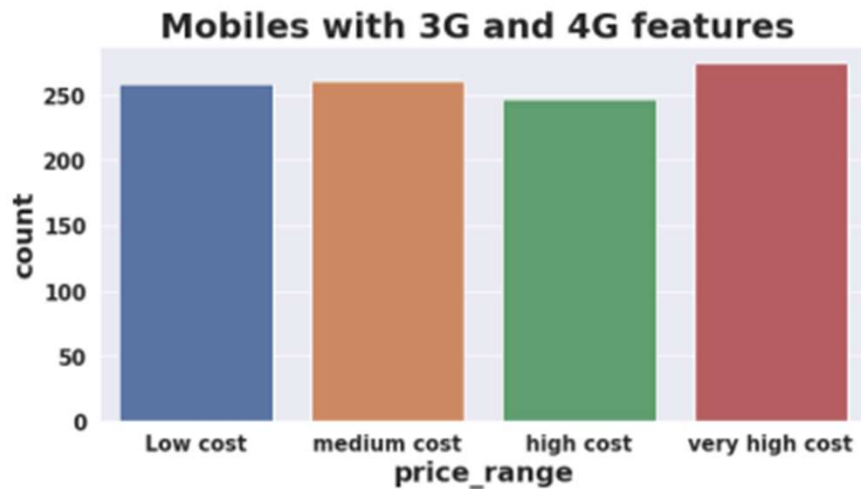
**Bivariate and Multivariate Analysis:** Different trends of price range v/s other features





# EDA (Exploratory Data Analysis):

## Bivariate and Multivariate Analysis:



- ❖ 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.

# 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.
- Decision Tree
- Random Forest classifier
- Gradient Boosting Classifier
- K-nearest Neighbor classifier
- XG Boost Classifier
- Support Vector Machine(SVM)

# Evaluation of models:

Algorithms	Training Set		Test set	
	Accuracy score (%)	Recall (%)	Accuracy Score	Recall (avg of all 4 classes)
Decision Tree	100	100	84	83.75
Decision Tree(Hyperparameter Tuning)	97.62	97.5	85.13	84.75
Random Forest	100	100	88.6	88.5
Random Forest ( HyperParameter Tuning)	100	100	89.81	89.5
Gradient Boosting	100	100	90.02	90
Gradient Boosting(HyperParameter Tuning)	100	100	90.42	90.5
KNN	75.86	76	59.47	59.25
KNN(HyperParameter Tuning)	76.61	76.75	70.26	69.75
XG-Boost	98.98	98.75	90.22	90
XG-Boost (HyperParameter Tuning)	100	100	92.46	92.25
SVM	98.57	98.5	89.81	89.75
SVM(HyperParameter Tuning)	98.3	98.5	97.96	98

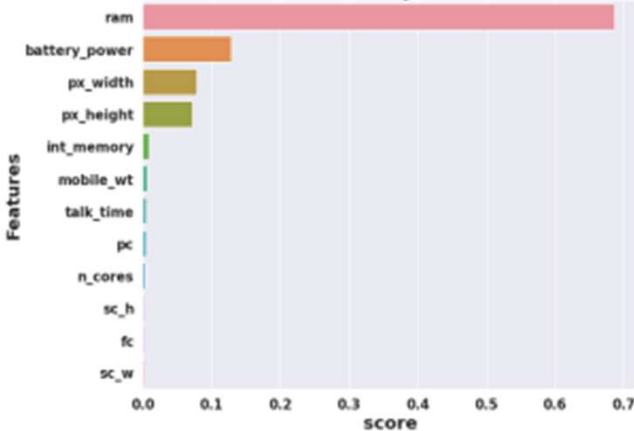
- ❖ Best model came out to be SVM after hyperparameter tuning.
- ❖ XG boost (Hyper-parameter Tuned) can be considered as the second most good model.
- ❖ NN performed very worst.

# Feature importance's :



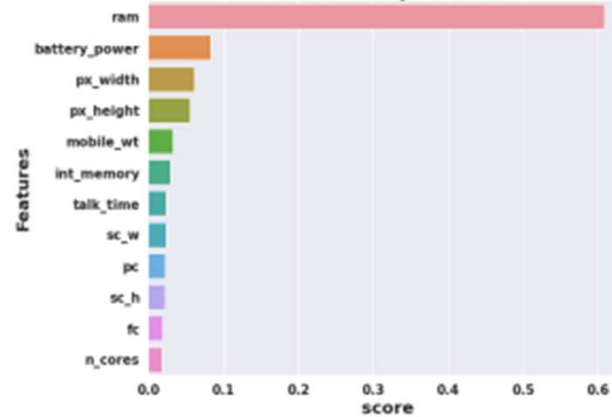
Decision Tree

Feature Importance



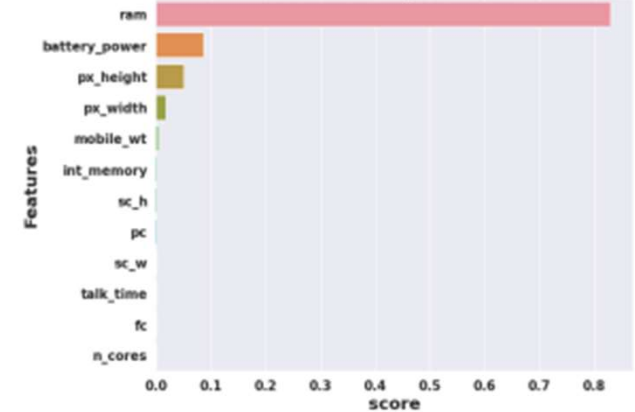
Random Forest

Feature Importance



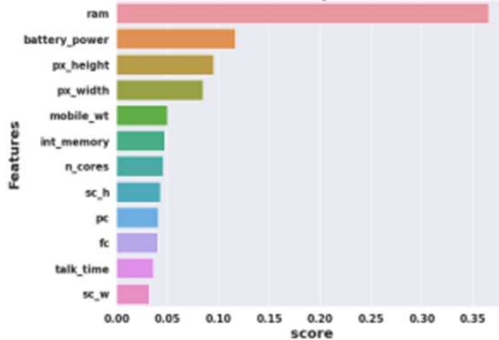
Gradient Boost

Feature Importance

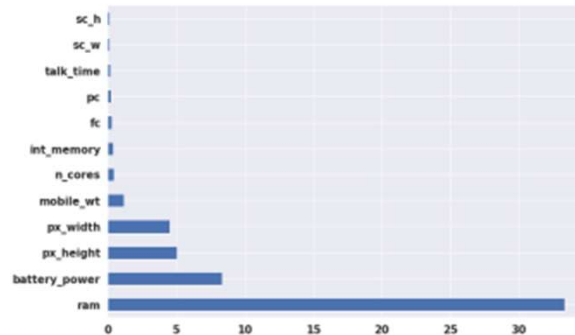


XG boost

Feature Importance



SVM

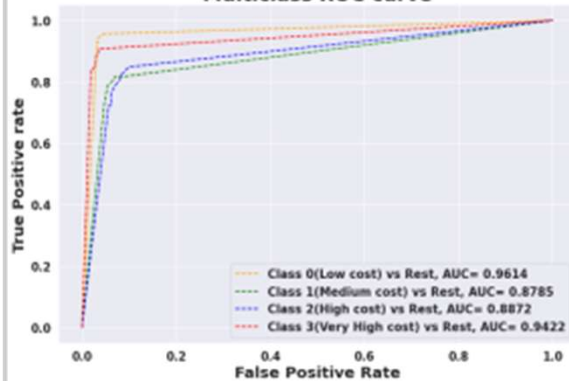


RAM, Battery Power, Pixel height and weight contributed the most in predicting the price range.

# AUC ROC curves:

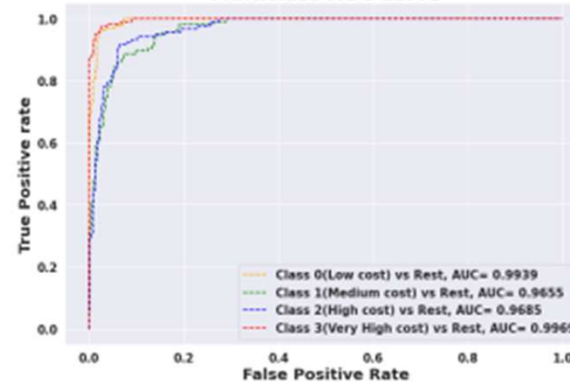
Decision Tree

Multiclass ROC curve



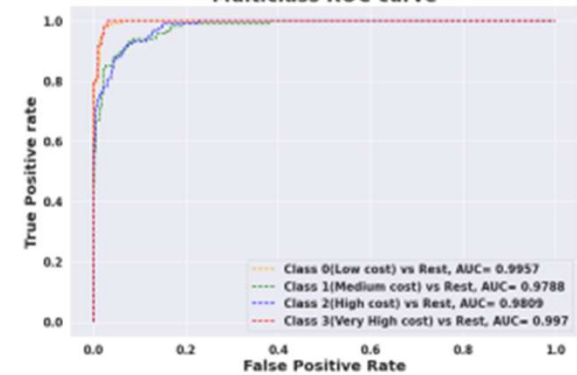
Random Forest

Multiclass ROC curve



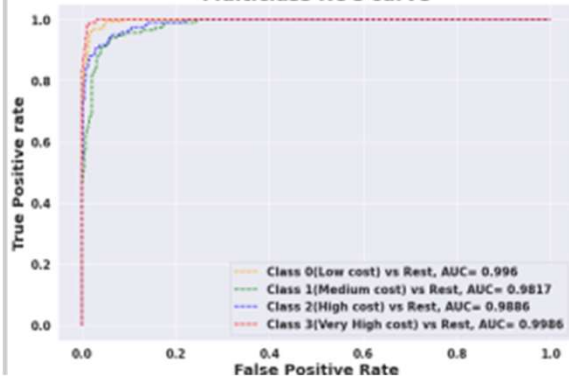
Gradient Boost

Multiclass ROC curve



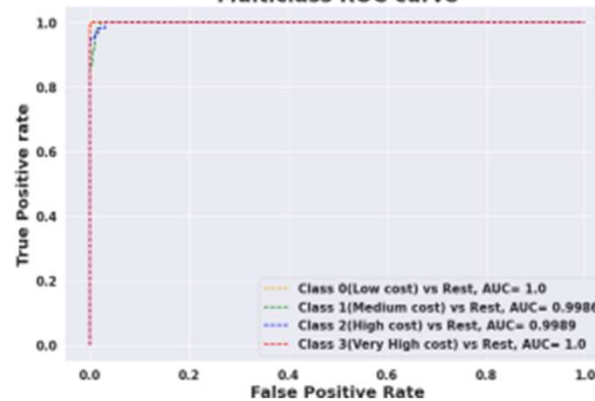
XG boost

Multiclass ROC curve



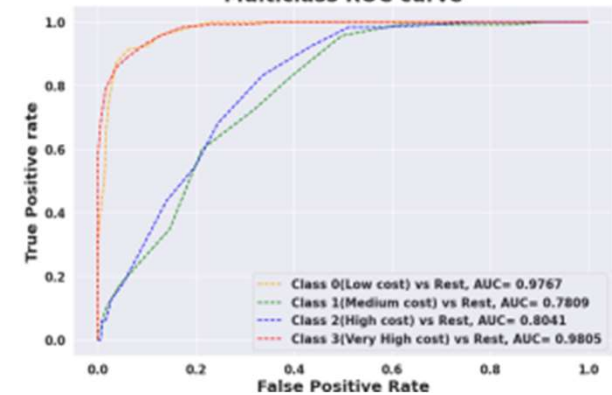
SVM

Multiclass ROC curve



KNN

Multiclass ROC curve



# 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 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.

**Thank You**