Capstone Project-3

Mobile Price Range Prediction

Supervised Machine Learning (Classification)

TEAM MEMBERS

Adi Ingrole

Mandar Khatavkar

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:

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.
- **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.
- **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 Mismatch values in data.

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

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

```
180
```

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

```
2
```

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

```
# Checking How many observations having sc_w value as 0.
mobile_data[mobile_data['sc_w']==0].shape[0]
```

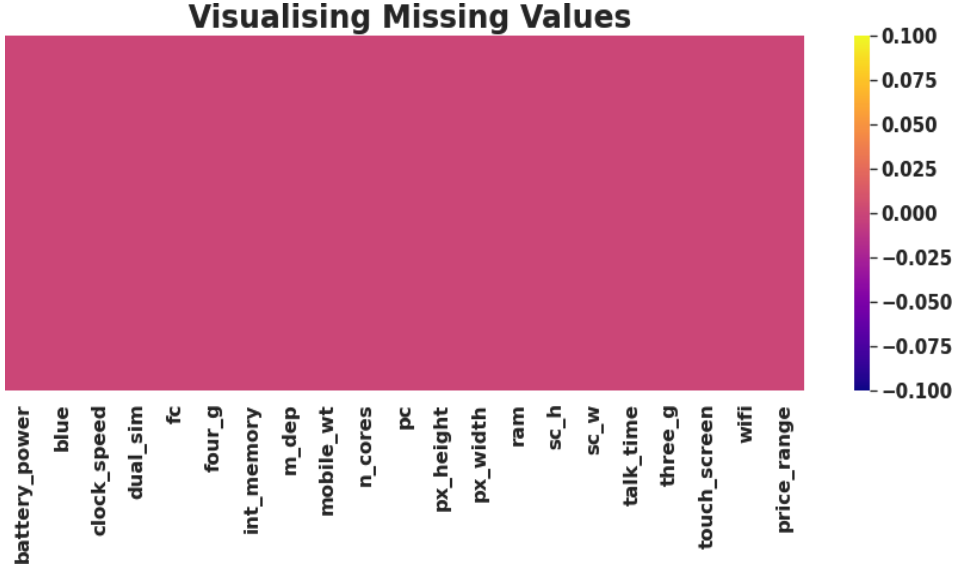
```
0
```

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

Data Wrangling :

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

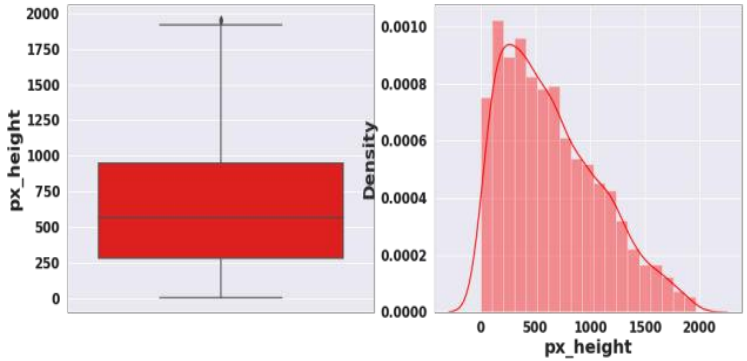
```
 We have 0 duplicate values in dataset.
```

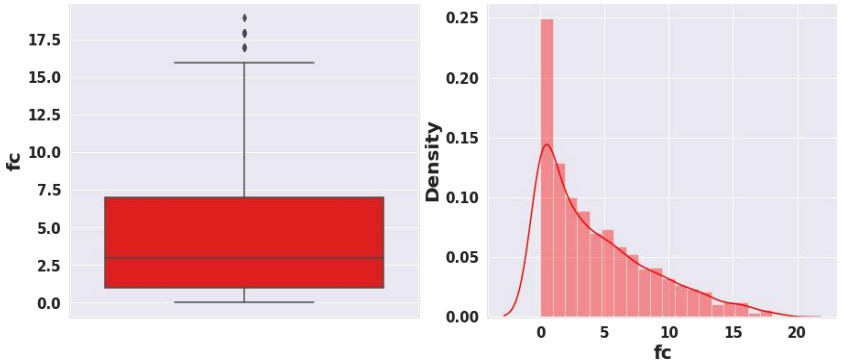


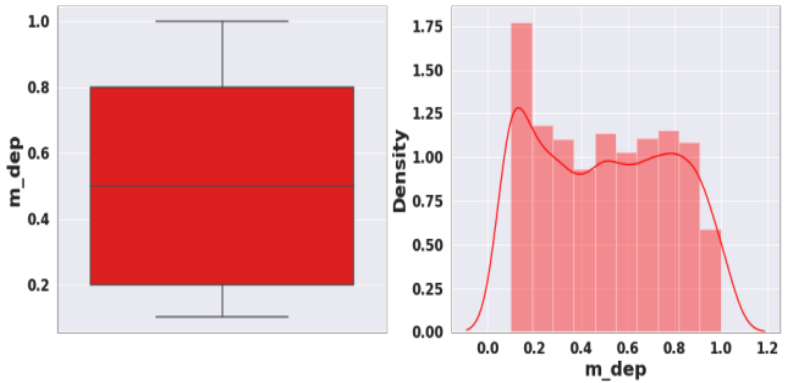
- Zero Missing values after handling mismatch from the data.
- 0 duplicates.

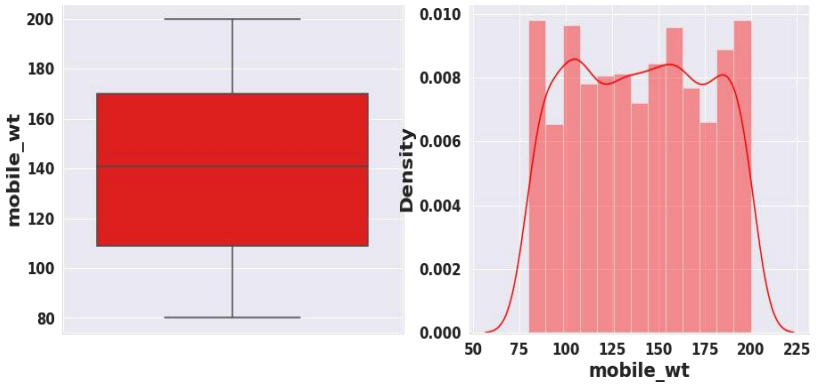
Data Wrangling :

- Checking outliers and Distribution of numerical variables

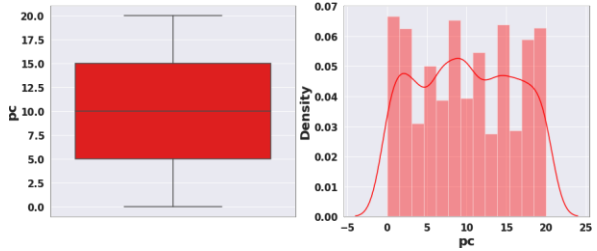


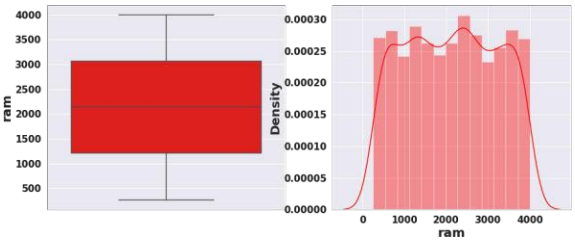


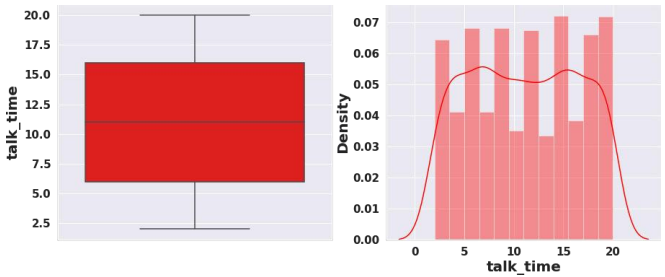


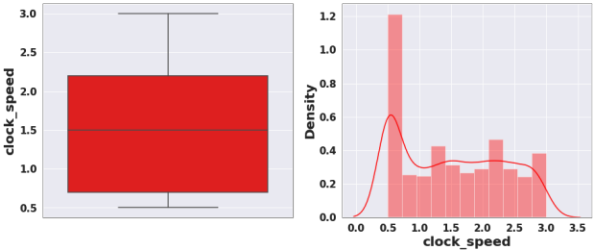


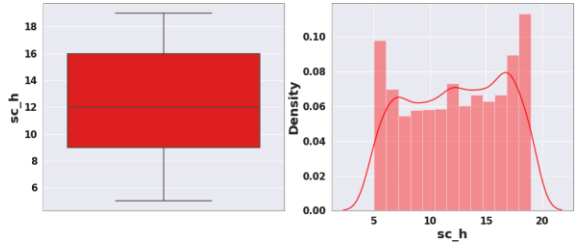
Data Wrangling :

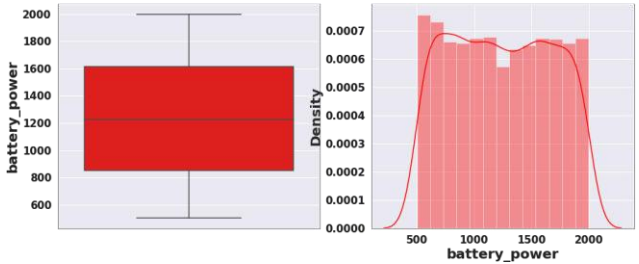


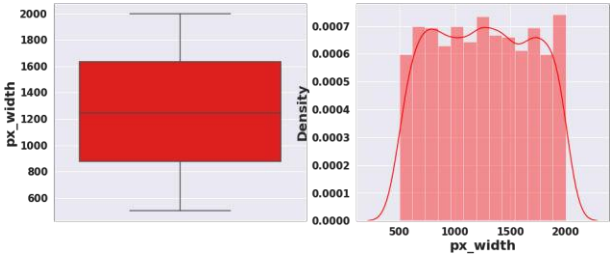


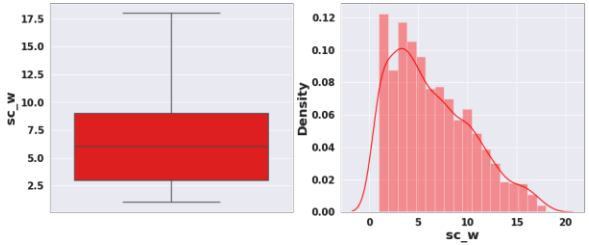


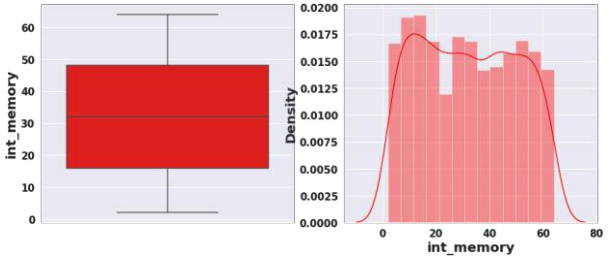






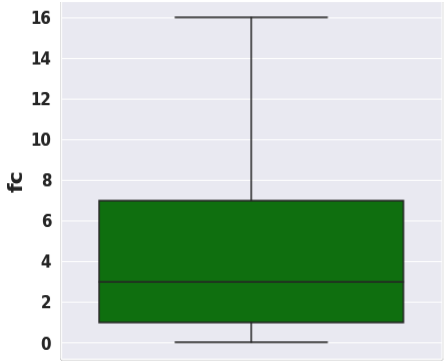


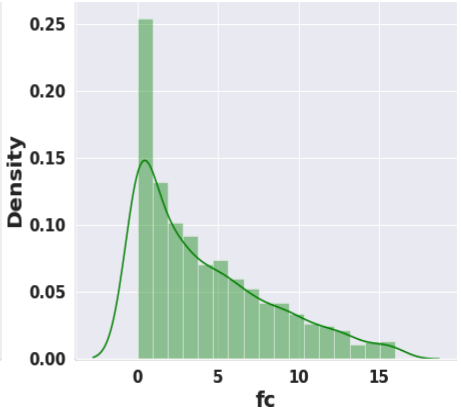


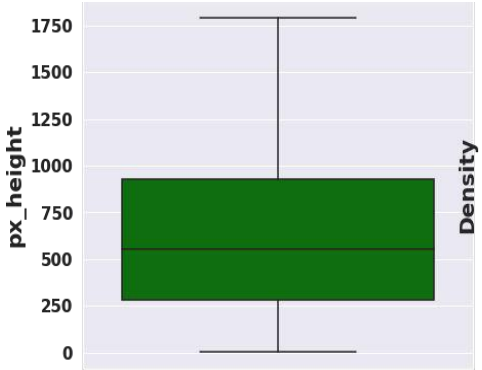


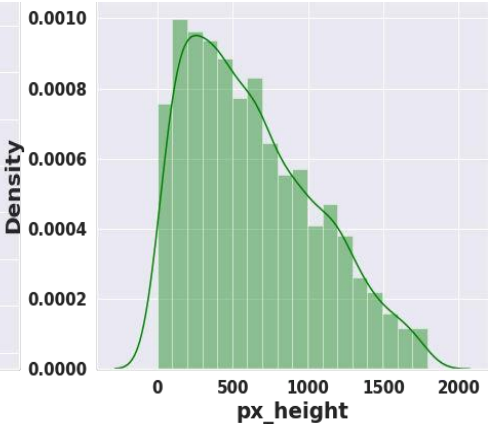
Data Wrangling :

- After removal of outliers



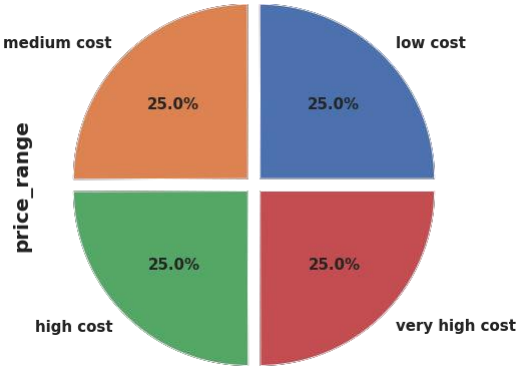


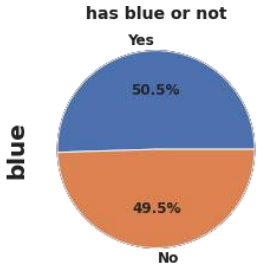


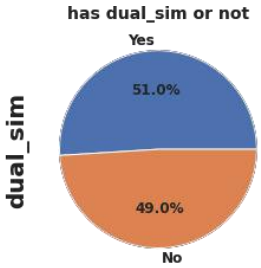


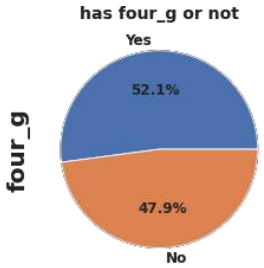
Exploratory Data Analysis

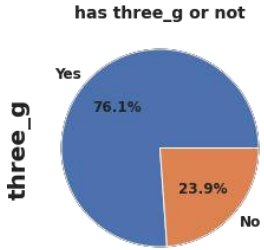
Univariate Analysis

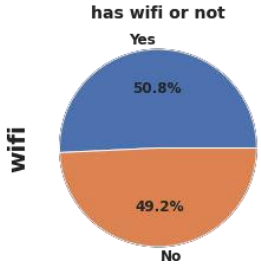


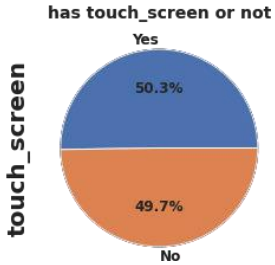








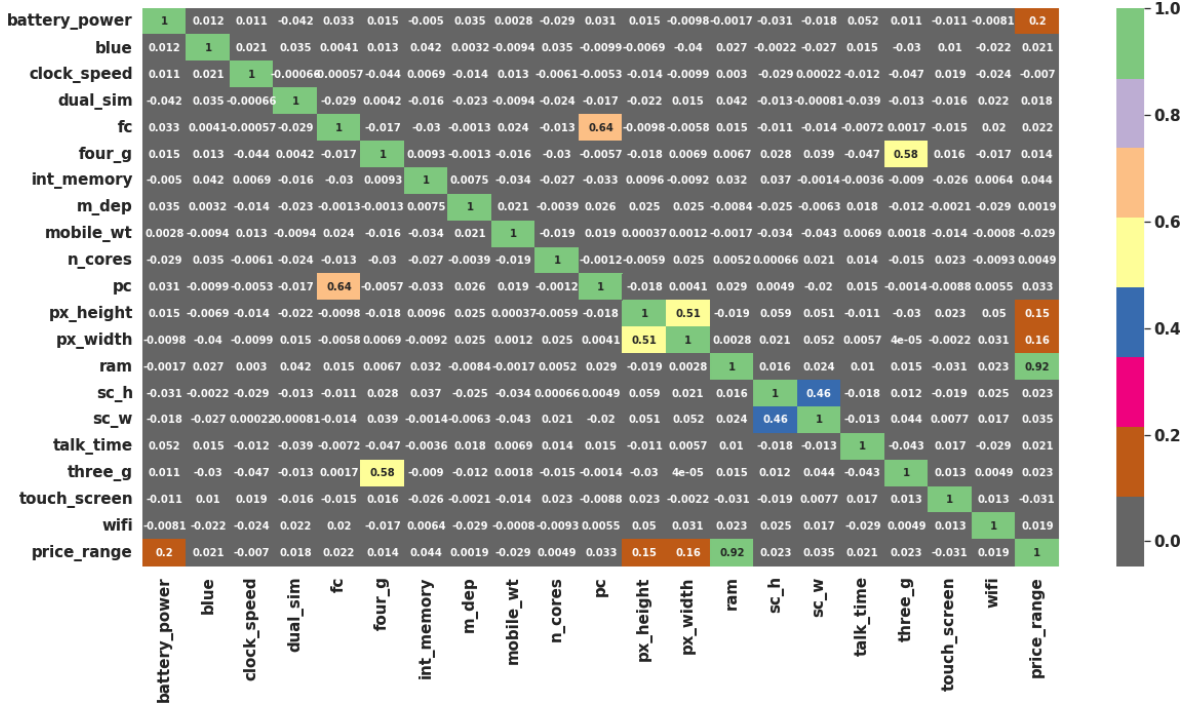




- Our target variable has equal number of observations in each category. Target variable is equally distributed.
- Percentage Distribution of Mobiles having bluetooth, dual sim, 4G,wifi and touch screen arealmost 50 %.
- Very few mobiles(23.8%) do not have 3G .

Bivariate and Multivariate Analysis

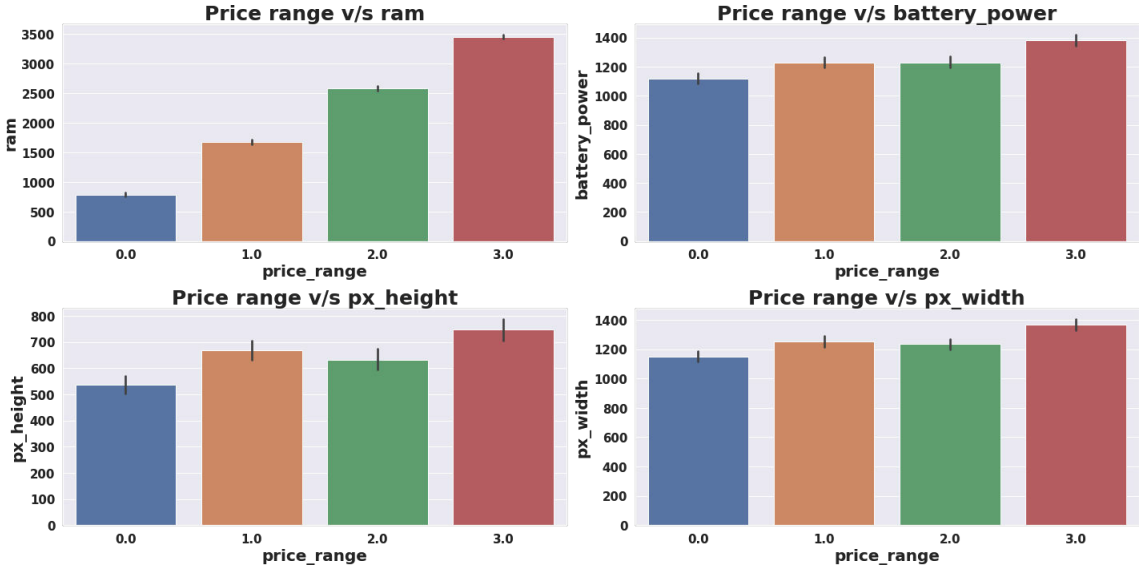
Correlation of independent variable with target variable.



- RAM has strong positive correlation withthe 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. Generallymobiles having high prices comes with good battery power.
- Also px_height and px_width (Pixel Resolution Height and width) are positivelycorrelated. Generally High price range mobiles have good resolutions.
- Four_g and Three_g are highly positvely correlated. Nowdays most of the smart mobiles has both type of options. This could be the reason that they are correlated.

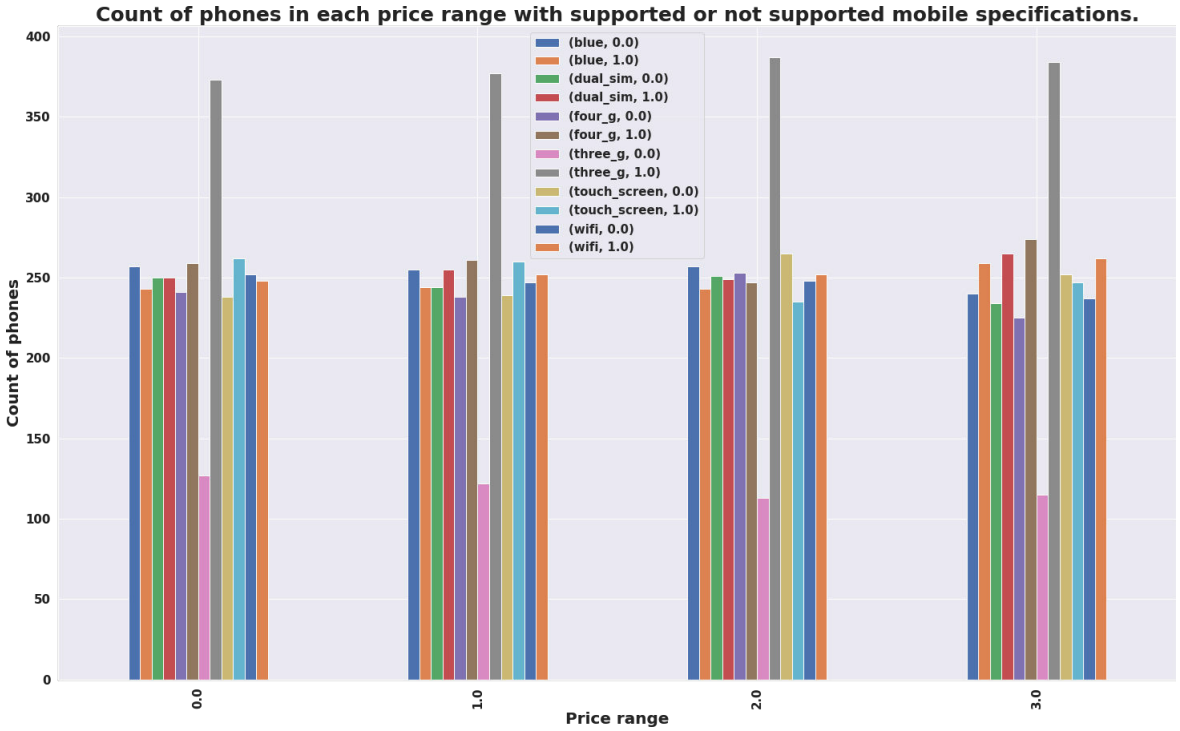
- primary camera i.e pc and front camera fc are positively correlated.
- sc_h and sc_w are positively correlated.

Bivariate and Multivariate Analysis:



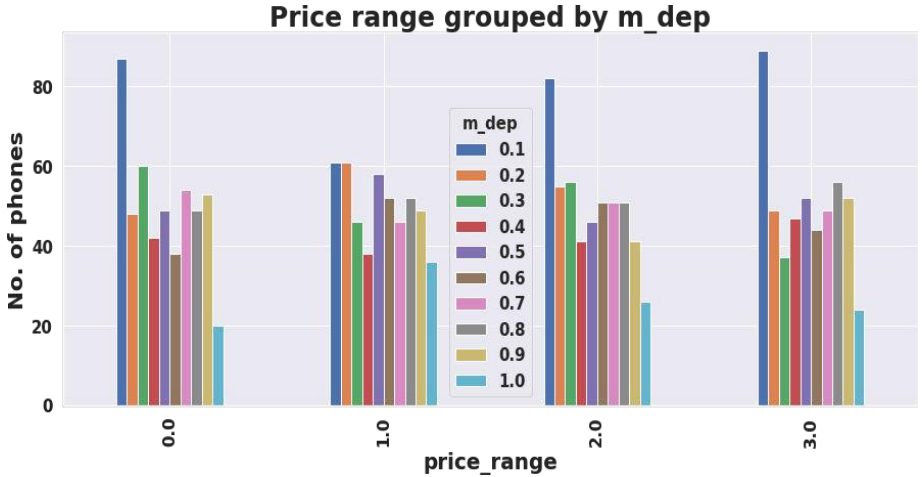
- Mobiles having RAM more than 3000MB falls under Very high cost category. As RAMincreases price range also increases.
- Mobiles having RAM less than 1000 MB falls under low cost category.
- Mobiles with battery power more than 1300 mAh has very high cost. And Mobileswith battery power between 1200 and 1300 mAh falls under medium and high cost category.
- Mobiles with more than 700 pixel height and width more than 1300 has very highcost.

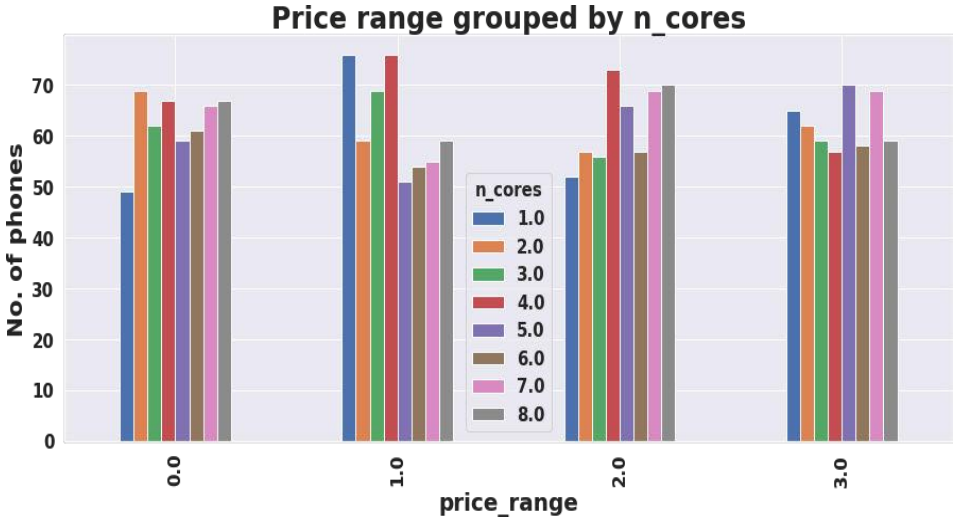
Bivariate and Multivariate Analysis:



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

Bivariate and Multivariate Analysis:

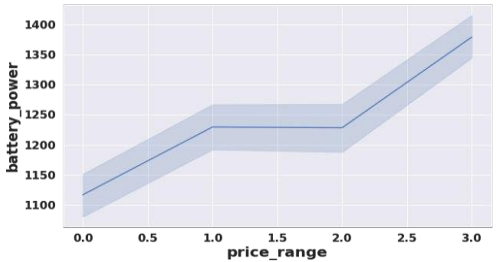


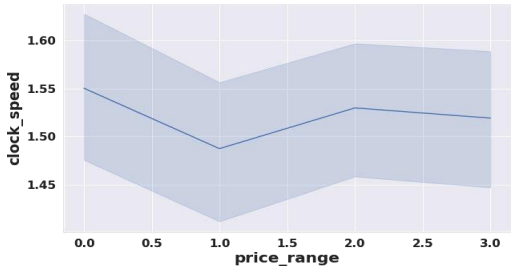


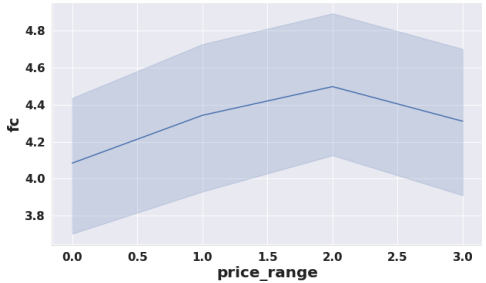
- There are very few mobiles in price range 0 and 1 with lesser no of cores.
- Most of the mobiles in price range 2 and 3 are with high no of cores.
- Number of phones with less thickness is high and count of phones with high thickness is low.

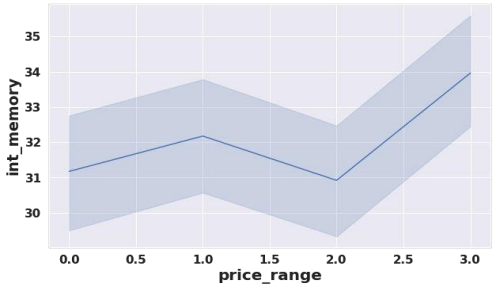
Bivariate and Multivariate Analysis:

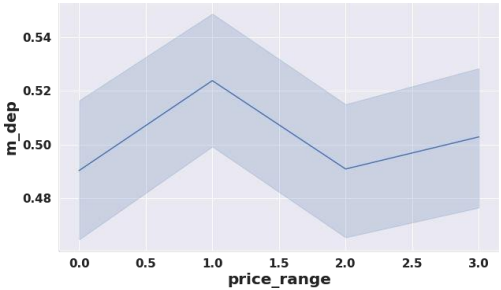
Different trends of price range v/s other features

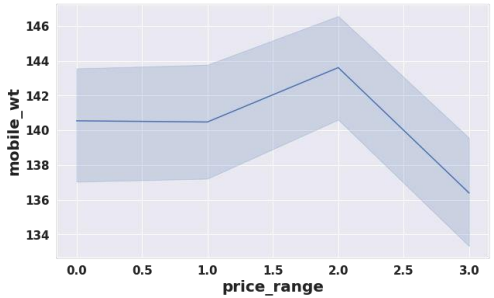


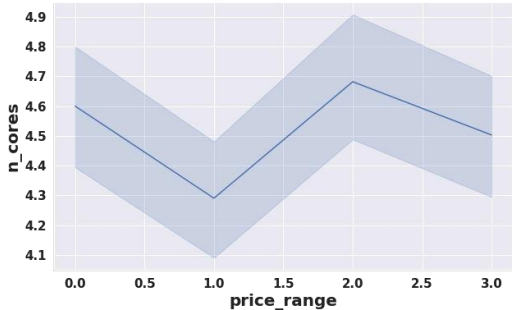


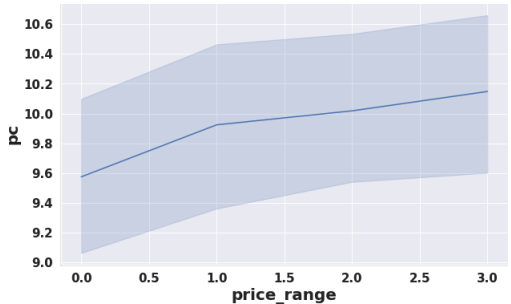


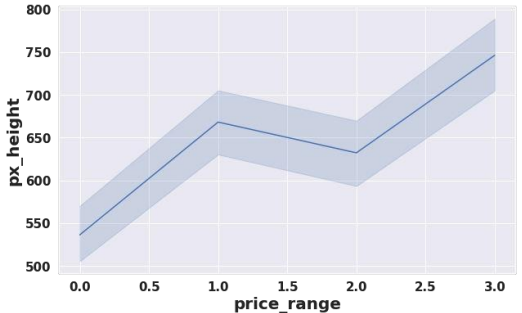




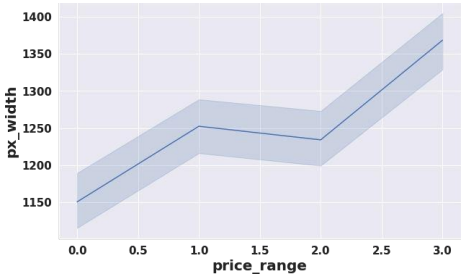


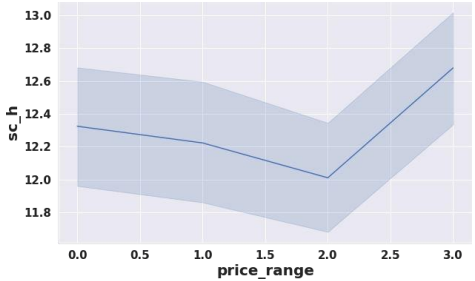


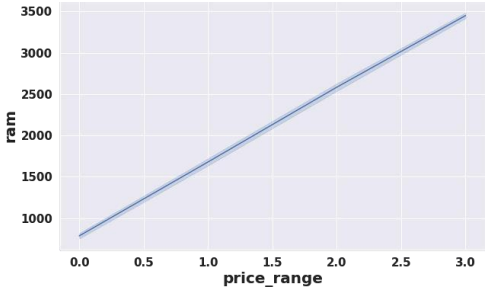


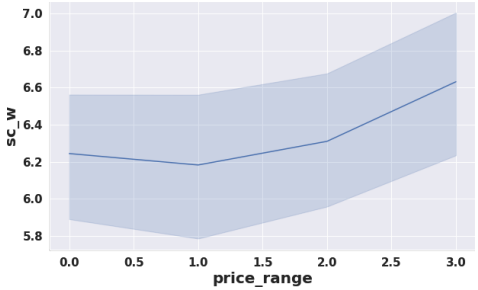


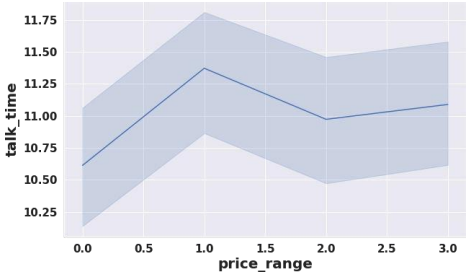
Bivariate and Multivariate Analysis:



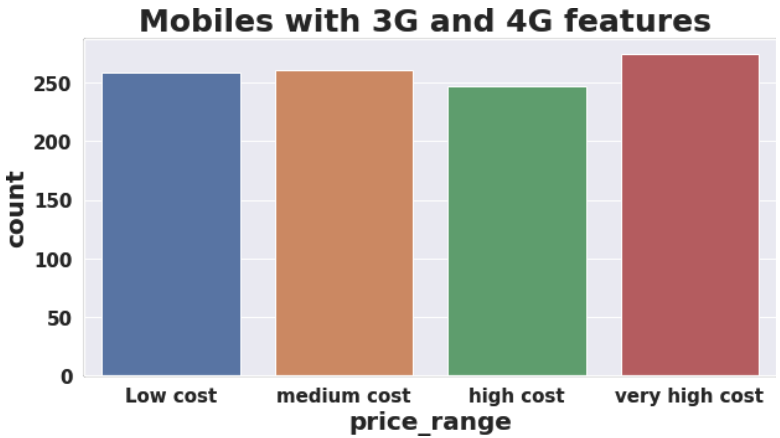


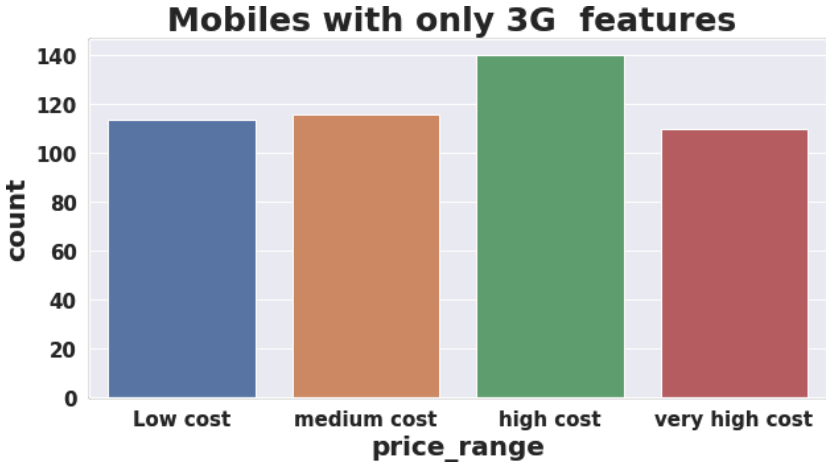






Bivariate and Multivariate Analysis:





- 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 remaining75% 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)

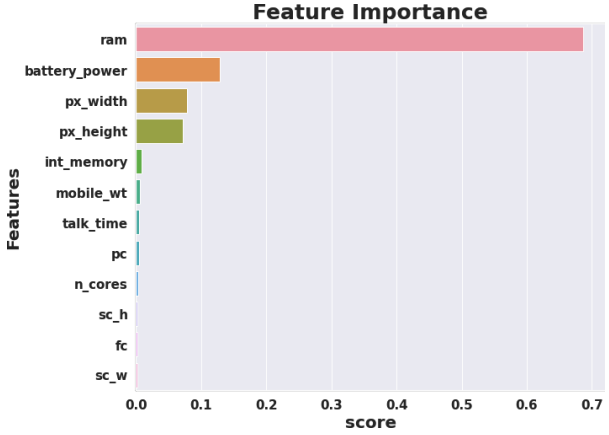
Evaluation of models:

Algorithms	Training Set		Test set	
	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

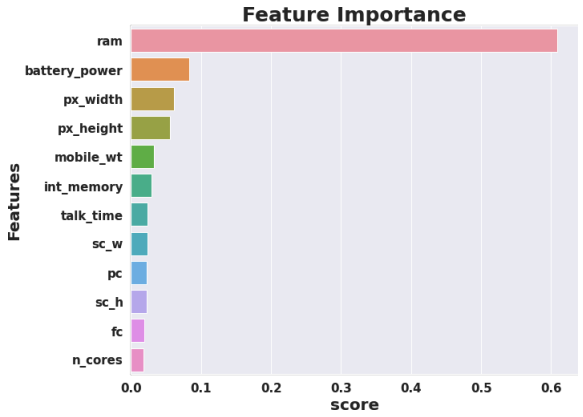
- Best model came out to be SVM after hyper-parameter tuning.
- XG boost (Hyper-parameter Tuned) can be considered as the second most good model.
- KNN performed very worst.

Features Importance:

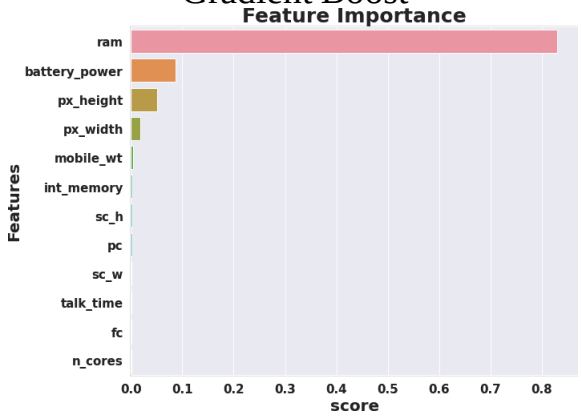
Decision Tree



Random Forest

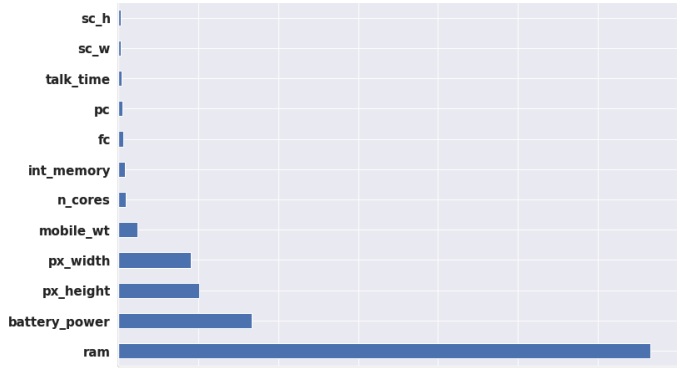


Gradient Boost



XG boost

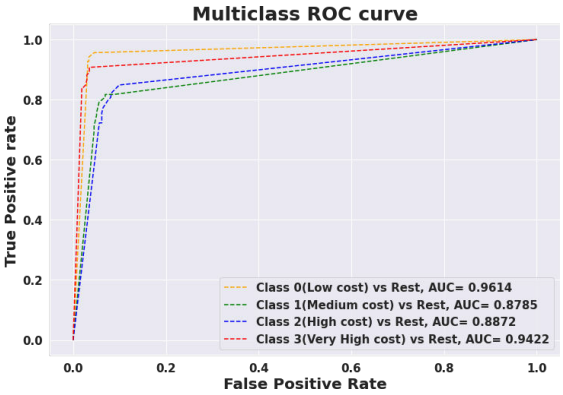
SVM



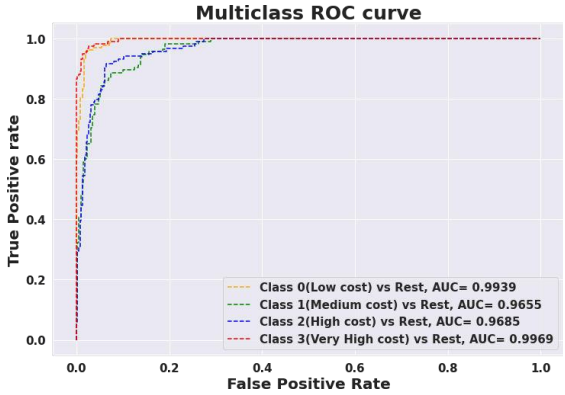
RAM, Battery Power, Pixel height and weight contributed the most in predicting the price range.

ROC curve:

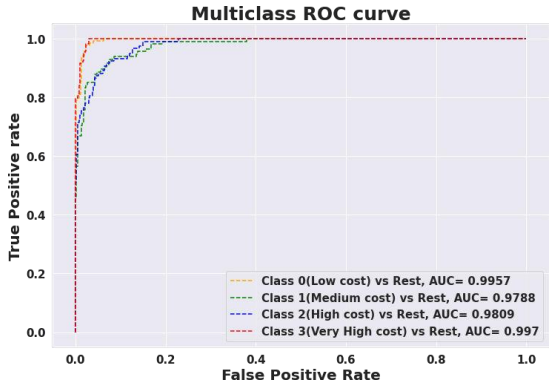
Decision Tree



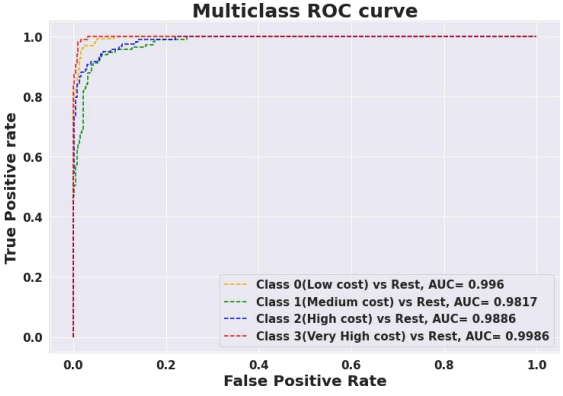
Random Forest



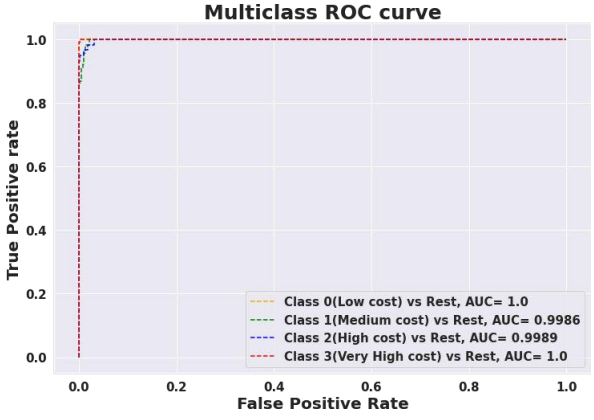
Gradient Boost



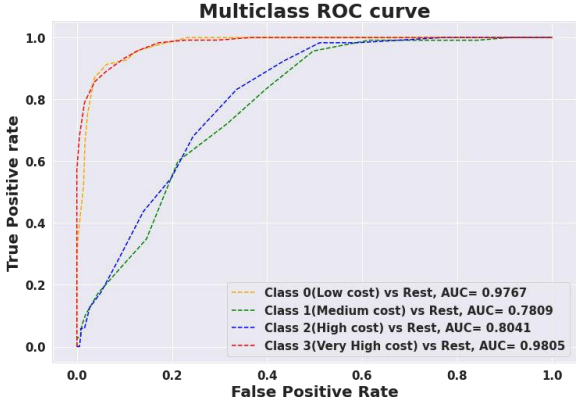
XG boost



SVM



KNN



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 usingChi square statistic.

- Implemented various classification algorithms, out of which the SVM(Support vector machine) algorithmgave 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 tuningwith 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_widthcontributed the most while predicting the price range.

Thank You!

Adi Ingrole
Mandar Khatavkar