Capstone project - 3

Mobile Price Range Prediction

Supervised Machine Learning (Classification)

BY

Nikhil Khushal Nimje

Introduction

Mobile phones are the best selling electronic devices as people keep updating their cell phones whenever they find new features in a new device. Thousands of mobiles are sold daily, in such a situation it is a very difficult task for someone who is planning to set up their own mobile phone business to decide what the price of the mobile should be.

The price of a product is the most important attribute of marketing that product. One of those products where price matters a lot is a smartphone because it comes with a lot of features so that a company thinks a lot about how to price this mobile which can justify the features and also cover the marketing and manufacturing costs of the mobile.

In the section below, I will introduce you to a machine learning project on a mobile price classification model where I will train a model to classify the price range of mobiles using Python.

Points to be Discuss

- Methodology
- Data Summary
- Data Collection
- Data Wrangling
- EDA
- Model Building
- Evaluation Of Model
- Feature Importance
- conclusion

Methodology

- We will proceed with reading the data, and then perform data analysis.
- The practice of examining data using analytical or statistical methods in order to identify meaningful information is known as data analysis.
- After data analysis, we will find out the data distribution and data types.
- We will train 6 classification algorithms to predict the output. We will also compare the outputs. Let us get started with the project implementation.

Data Summary

- We have a data_mobile_price_range data for our analysis and model building.
- This data set contain total rows 2000 & total columns 21.
- This data set contain battery_power, blue, clock_speed, dual_sim, fc, four_g, int_memory, m_dep, mobile_wt, n_cores, pc, px_height, px_width, ram, sc_h, sc_w, talk_time, three_g, touch_screen, wifi, price_range, dtype='object’.
- 000 Total features=21

Data Collection

- 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.

Data Collection

- 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).

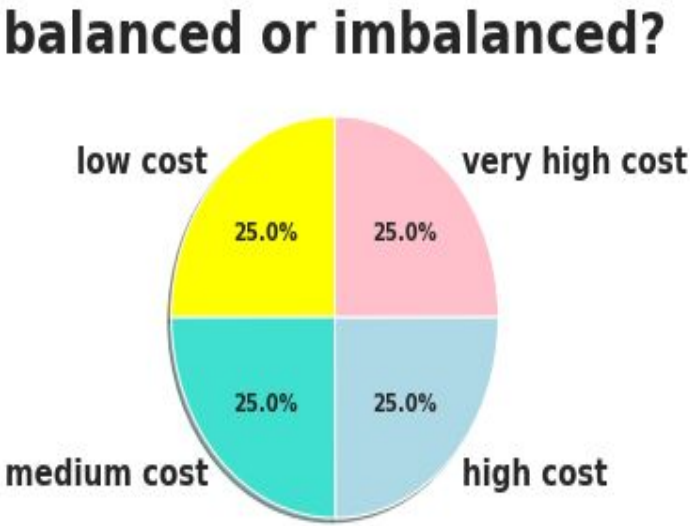
Data Wrangling

```
# null value count in training set
mobile_data.isna().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
dtype: int64
```

- Missing values are imputed using the K-Nearest Neighbors approach where a Euclidean distance is used to find the nearest neighbors.
- Zero Missing values after handling mismatch from the data.
- 0 duplicates

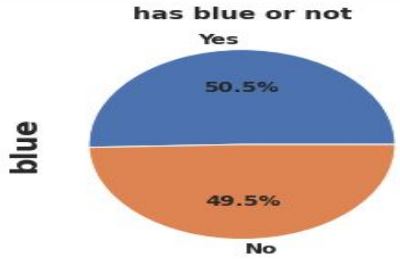
Data Wrangling

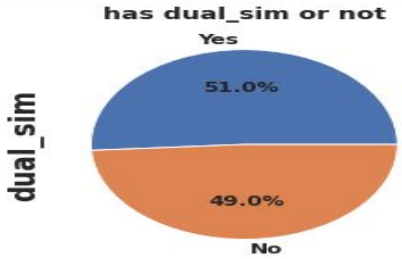


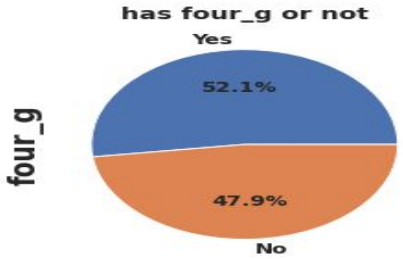
- The dataset provided is balanced as there is equal distribution of classes of price ranges.
- Thus we don’t have to worry about data imbalance and there is no need of oversampling or undersampling, which is good for us.

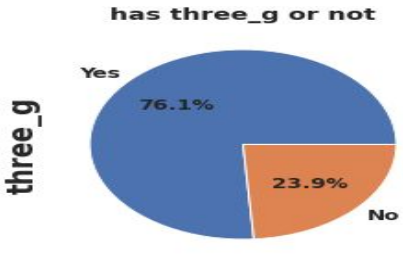
EDA

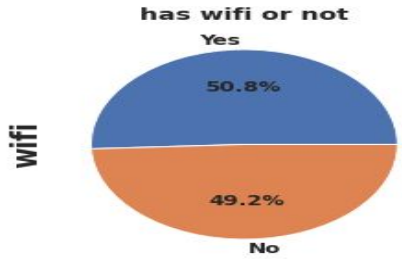
Univariate Analysis of Categorical columns.

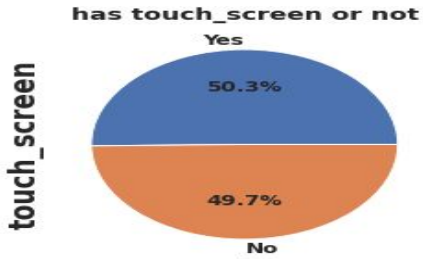








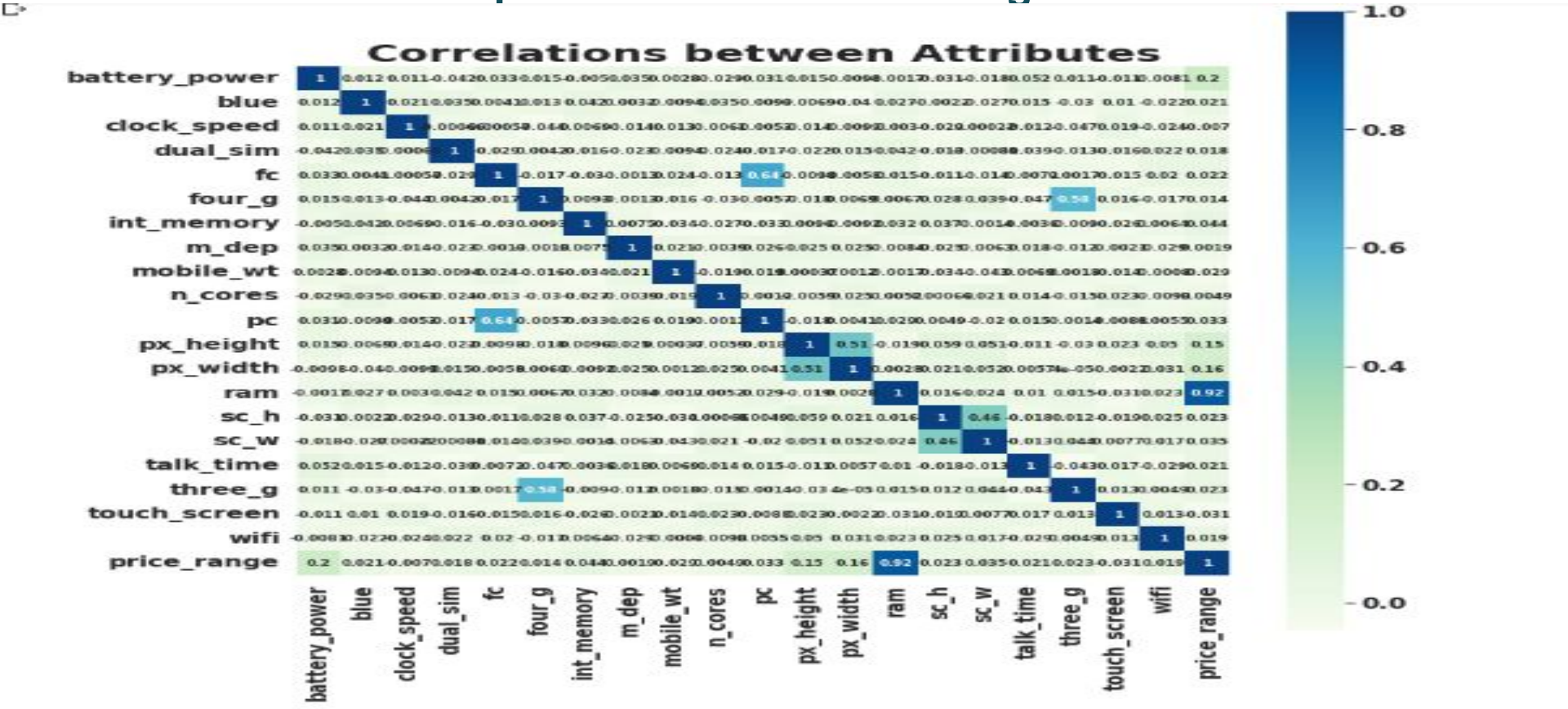




- Our target variable has equal number of observations in each category. Target variable is equally distributed.
- Percentage Distribution of Mobiles having bluetooth, dual sim, 4G,wifi and touch screen are almost 50 %.
- Very few mobiles(23.8%) do not have 3G .

EDA

Correlation of independent variable with target variable



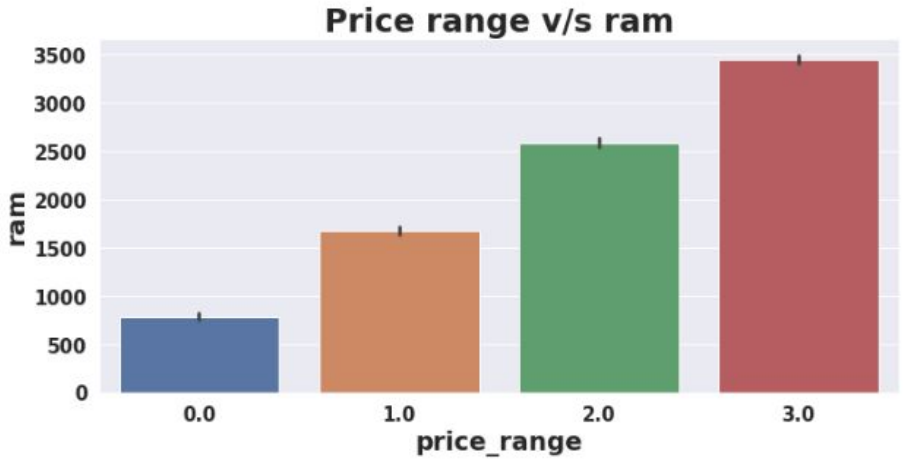
EDA

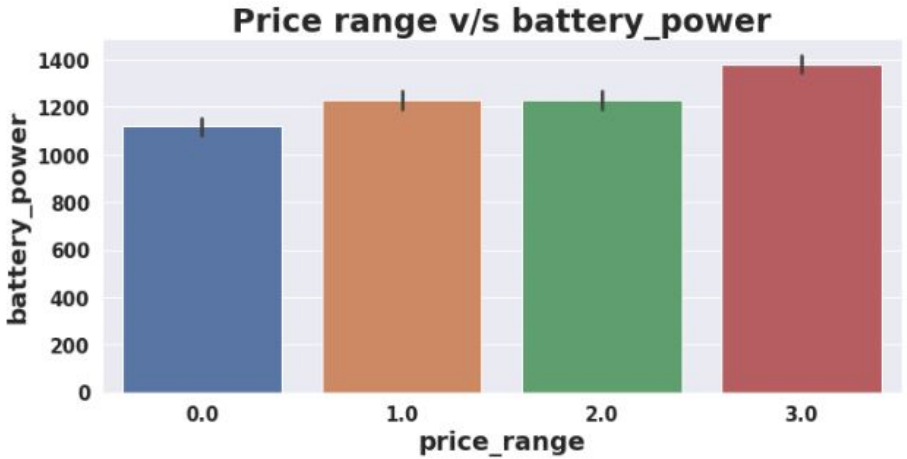
Correlation of independent variable with target variable

- 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.
- primary camera mega pixels and front Camera mega pixels have correlation (it make sense because both of them reflect technology level of resolution of the related phone model) but they do not effect price range.
- having 3G and 4G is somewhat correlated, Nowdays most of the smart mobiles has both type of options. This could be the reason that they are correlated.
- sc_h and sc_w are positively correlated.
- there is no highly correlated inputs in our dataset, so there is no multicollinearity problem.

EDA

Bivariate and Multivariate Analysis:

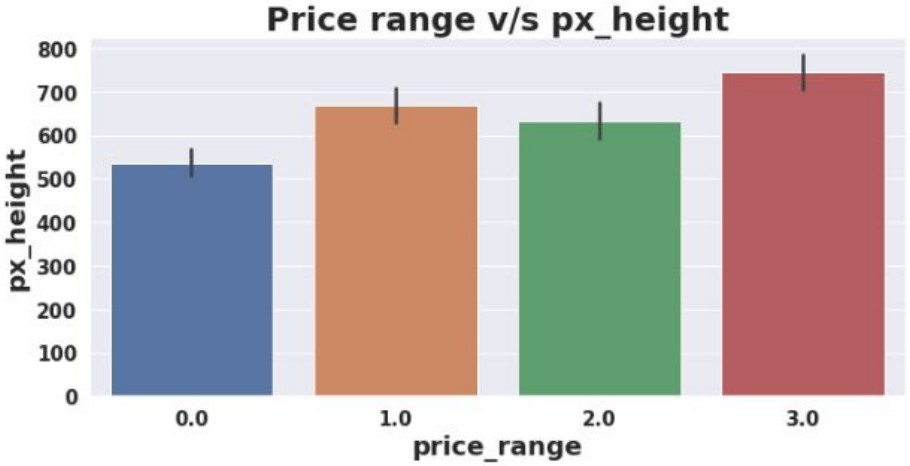


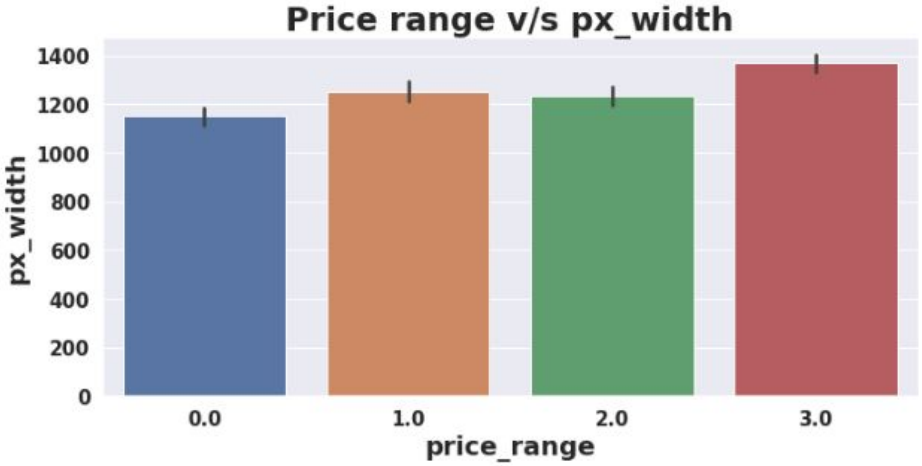


- Mobiles having RAM more than 3000MB falls under Very high cost category.As RAM increases price range also increases.
- Mobiles having RAM less than 1000 MB falls under low cost category.
- Mobiles with battery power more than 1300 mAh has very high cost. And Mobiles with battery power between 1200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

EDA

Bivariate and Multivariate Analysis:





- Mobiles having RAM more than 3000MB falls under Very high cost category.As RAM increases price range also increases.

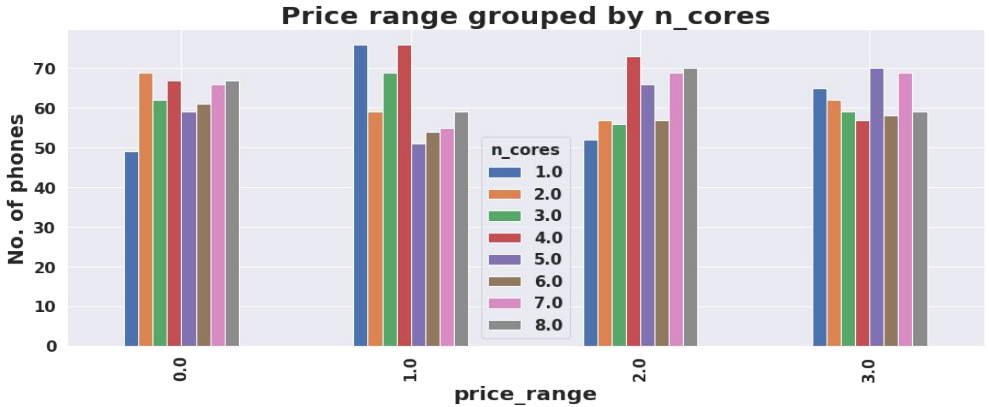
EDA

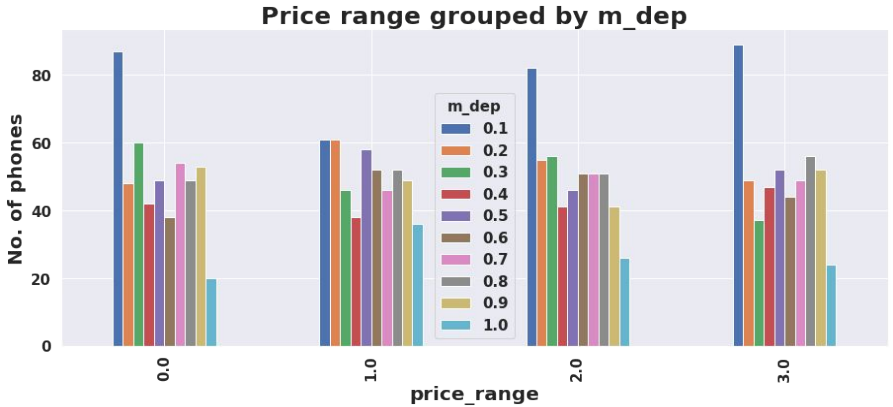
Bivariate and Multivariate Analysis:

- We can see that each price range category has equal number of mobiles phones having both supporting and non supporting specifications.

EDA

Bivariate and Multivariate Analysis:

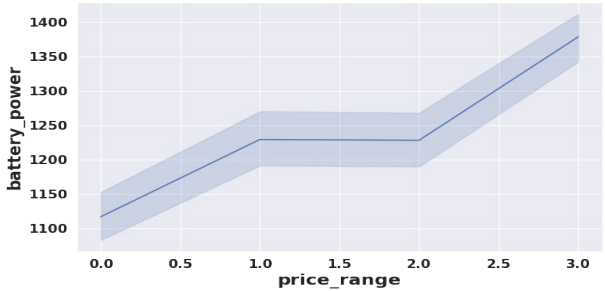


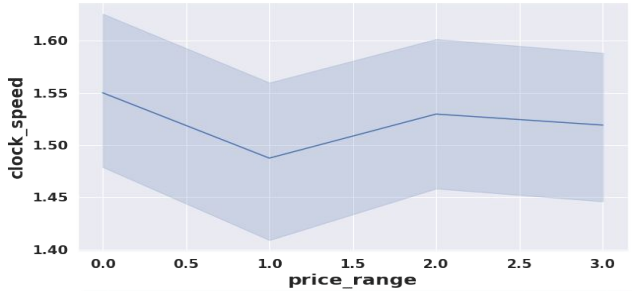


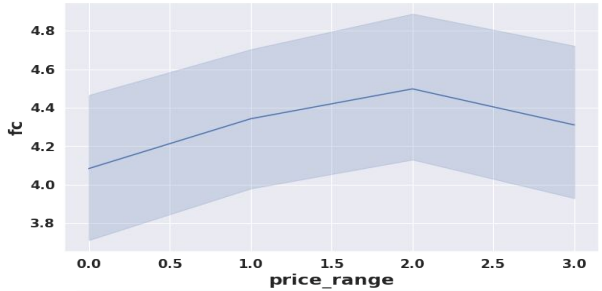
- Number of phones with less thickness is high and count of phones with high thickness is low.
- There are very few mobiles in price range 0 and 1 with lesser no of cores.
- Most of the mobiles in price range 2 and 3 are with high no of cores.

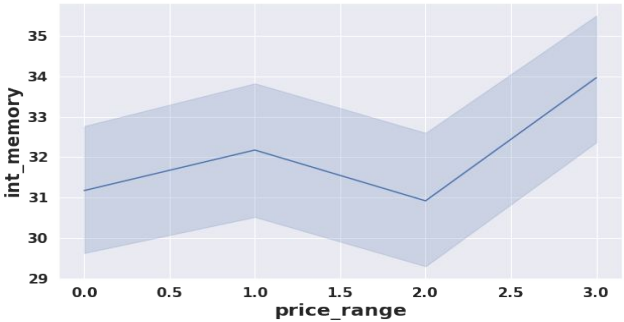
EDA

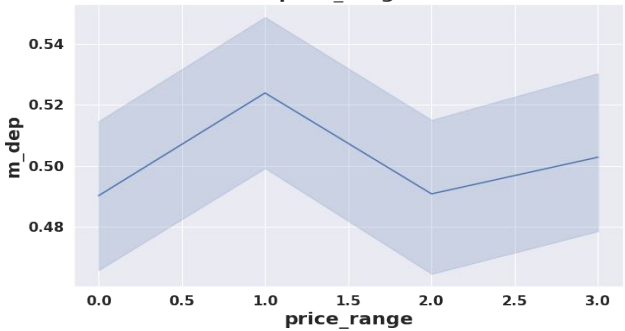
Different trends of price range v/s other

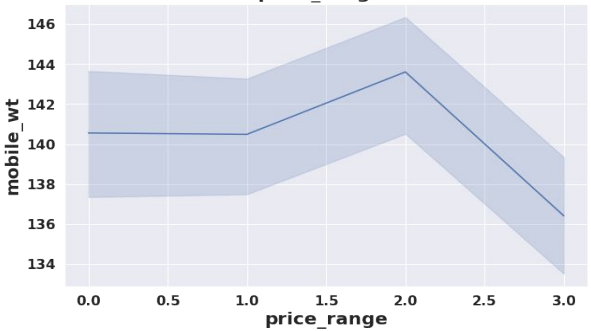


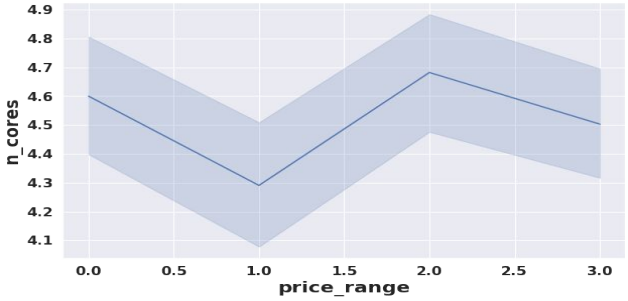


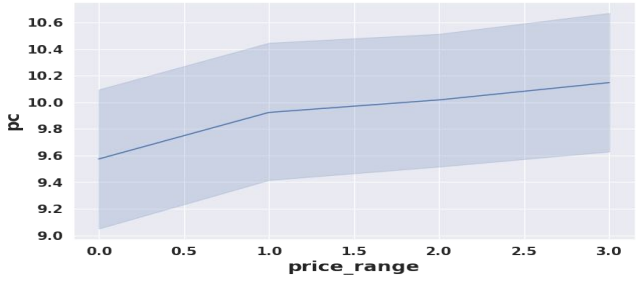


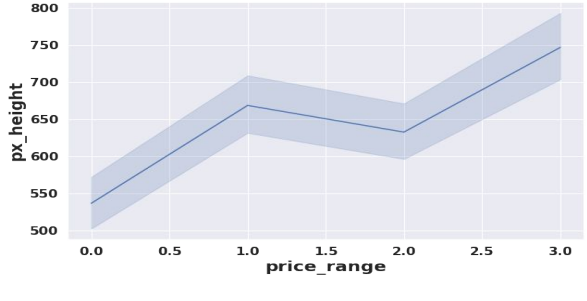






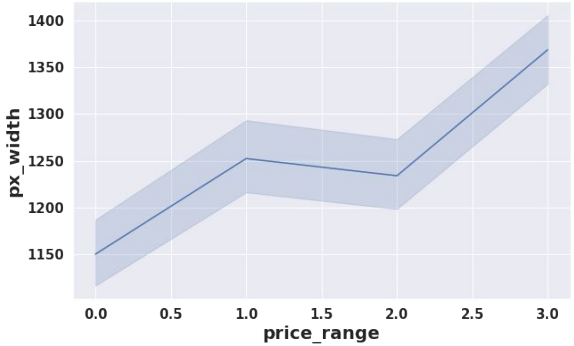


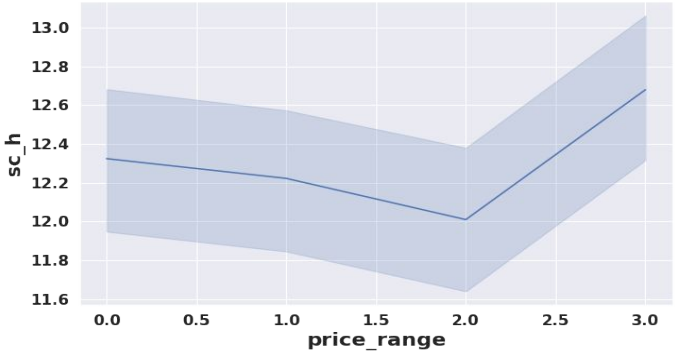


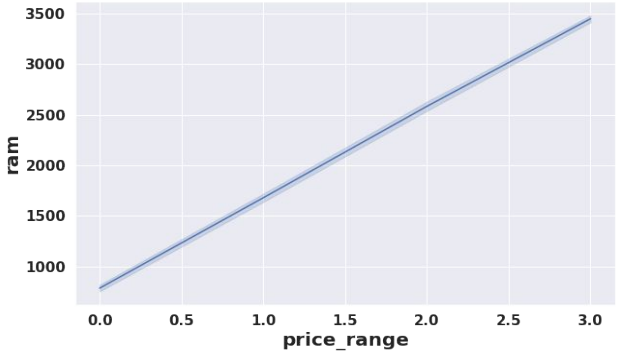


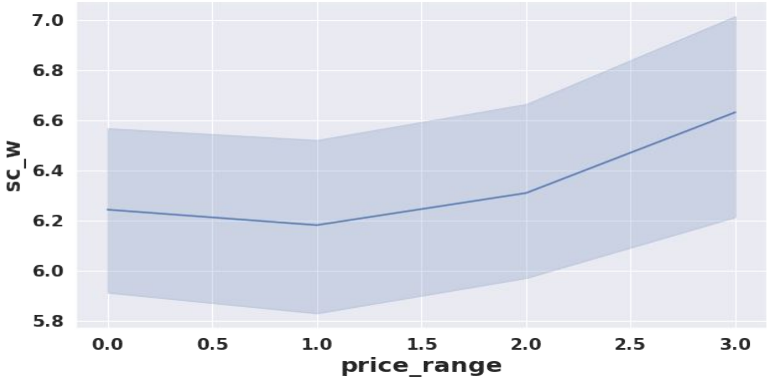
EDA

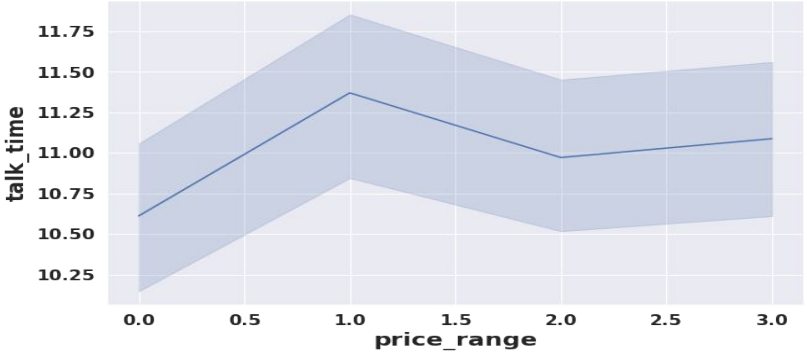
Different trends of price range v/s other features











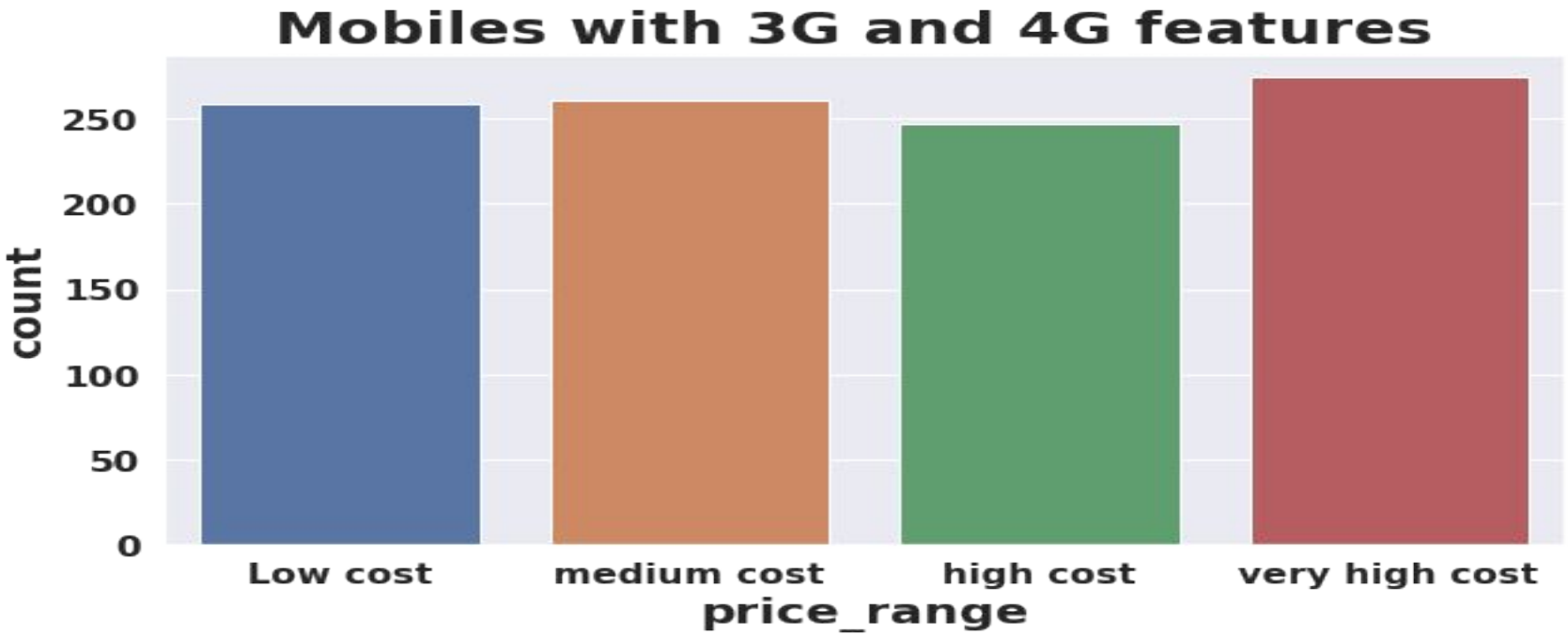
EDA

Different trends of price range v/s other features

- Mobiles having max screen height and width falls in very high price category. We can see in linechart of sc_width and sc_height from class 2 screen width and hieght starts increasing with price. Similar case is with px_height and px_width. As resolution of screen increases the price also increases RAM has clear relationship with price range we saw that in correlation matrix also.
- For class 1 and class2 battery power range is almost similar. As battery power increases price also increases whcih is quite obivious.
- Mobiles in very high price range(Class 3) has less weight compared to other classes.That means as weight of mobiles decrease price increases.

EDA

Mobiles with both 3G and 4G



- As we can see from low cost to very high cost mobiles have both features

Model Building

To predict the mobile phone prices, we are going to apply below algorithms respectively on the training and validation dataset. After that, we are going to choose the best model for our data set and create target values for test dataset.

- **Decision tree**
- **Random forest**
- **K-nearest Neighbour classifier**
- **Support Vector Machine(SVM)**
- **XG Boost Classifier**
- **Logistic regression**

1. As Decision tree,random forest and enssembles trees do not require Feature scaling as these are Tree based models. So we will be using X_train and X_test which are not scaled.
2. For K nearest Neighbors and SVM we will be usingseX_train_scaled and X_test_scaled. That is we we will use Standardised data. i.e. Scaled data. As these are distance based Algorithms.

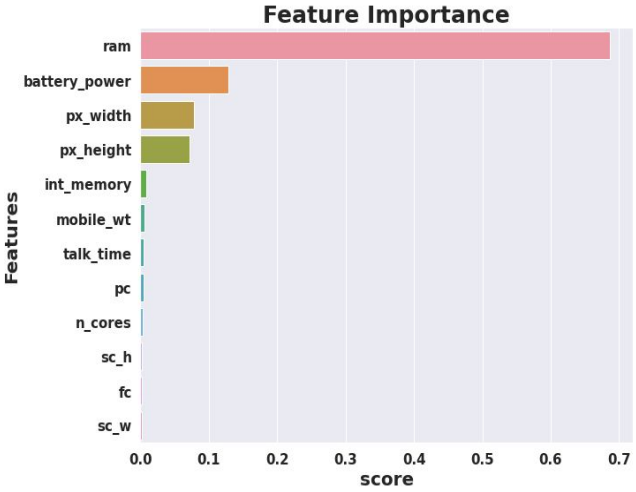
Evaluation of models:

Algorithms	Training Set		Test Set	
	Accuracy score (%)	Recall(%)	Accuracy score	Recall(avg of all 4 digits)
Decision tree	100	100	84	83.9
Decision tree (Hyper parameter tuning)	97.6	97	85.13	85
Random forest	100	100	88.5	88
Random forest (Hyper parameter tuning)	100	100	89.8	90
KNN classifier	75.8	76	59.4	59
KNN classifier (Hyper parameter tuning)	76.6	77	70.26	70
(SVM)	98.5	99	89.8	90
(SVM) (Hyper parameter tuning)	98.3	99	97.9	98
XG Boost	98.9	99	90.2	90
XG Boost (Hyper parameter tuning)	100	100	92.4	92
Logistic regression	63.15	63	64.9	65
Logistic regression (Hyper parameter tuning)	63.5	63	65.3	65

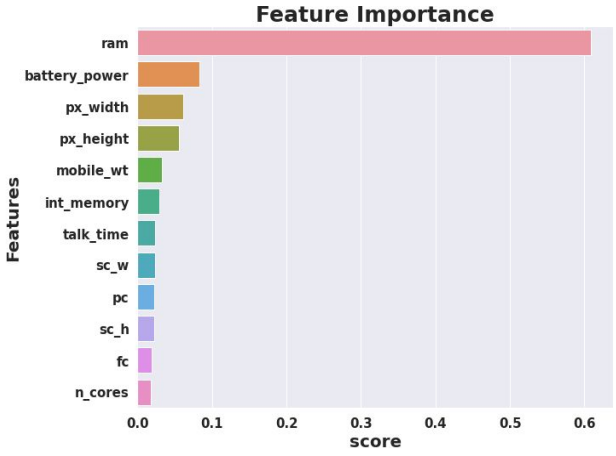
Evaluation of models:

- 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.

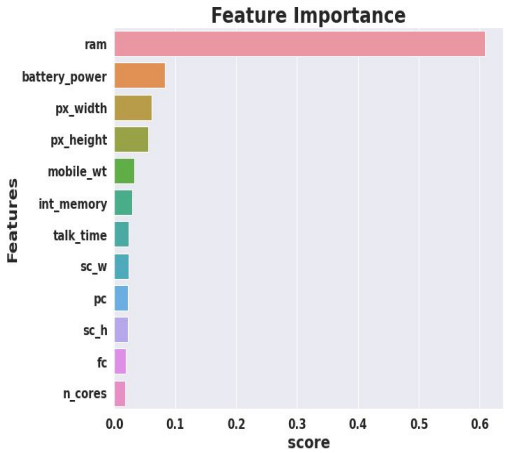
Feature Importance



Decision tree

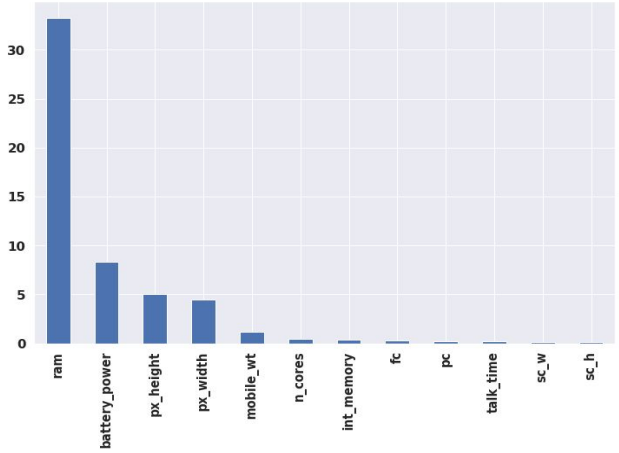


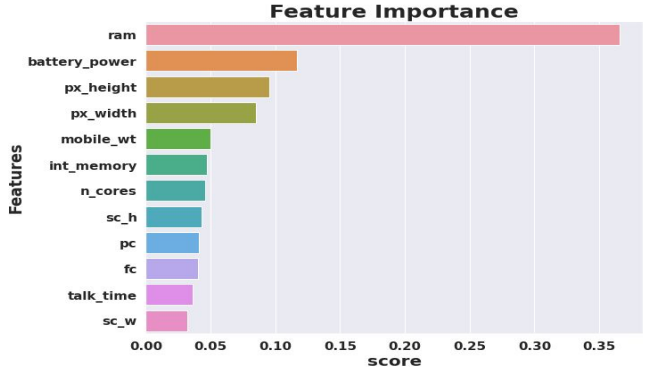
Random forest



KNN

SVM

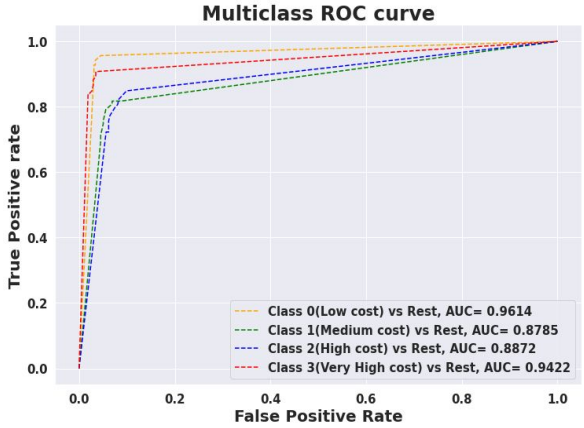




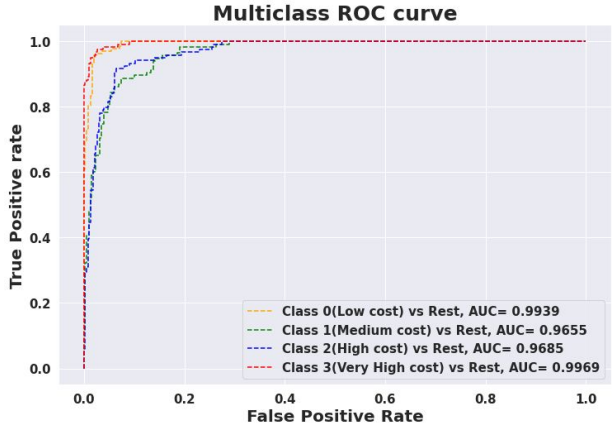
XG Boost classifier

AUC ROC curves:

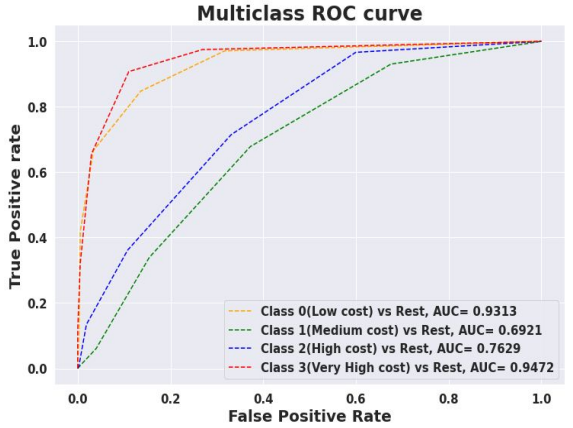
Decision tree



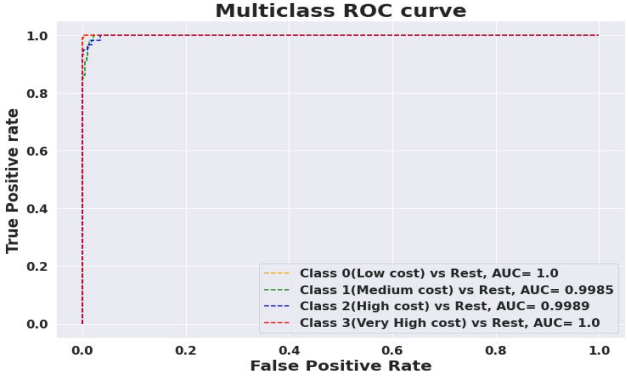
Random forest



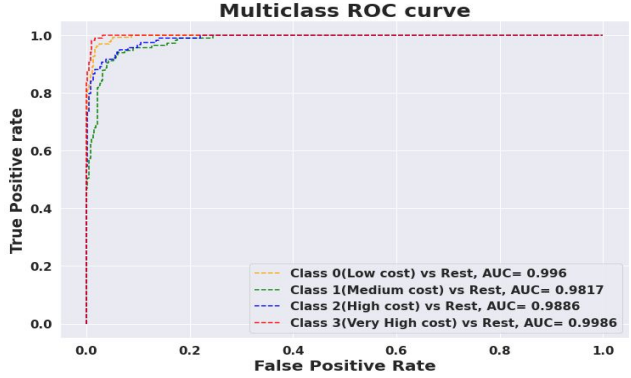
KNN



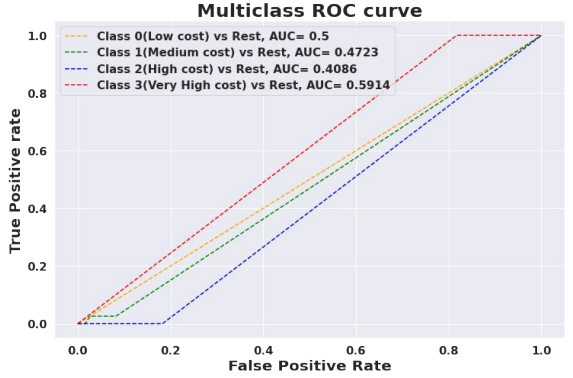
SVM



XG Boost classifier



Logistic Regression



Conclusions:

- We Started with Data understanding, data wrangling, basic EDA where we found the relationships, trends between price range and other independent variables.
- RAM, Battery Power, Pixel height and weight contributed the most in predicting the price range.
- 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.
- We selected the best features for predictive modeling by using K best feature selection method using Chi square statistic.
- KNN gave very worst model performance.
- 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.
- 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!