Project Description:

A mobile phone, cell phone, cellphone, or hand phone, sometimes shortened to simply mobile, cell or just phone, is a portable telephone that can make and receive calls over a radio frequency link while the user is moving within a telephone service area. It has become a part of human life. 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(eg:- 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

Problem Statement:

Develop a Supervised learning model using Classification algorithms to predict the price range of mobile phones in the ranges : 0(low cost), 1(medium cost),2(high cost) and 3(very high cost).
The objective is to find out some relation between features of a mobile phone(eg:- 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 Summary:

We have records of 2000 mobile phones with 20 columns/features. Each column represents the features of the mobile. We have zero null values.

Data Description:

The columns that we have in our dataset are:

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 megapixels
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 megapixels
Px_height - Pixel Resolution Height
Px_width - Pixel Resolution Width
Ram - Random Access Memory in MegaBytes
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
Touch_screen - Has touch screen or not
Wifi - Has wifi or not
Price_range - This is the target variable with values of 0(low cost), 1(medium cost),2(high cost) and 3(very high cost).

Steps followed towards a solution to our problem statement:

- **Exploratory Data Analysis:**
  It includes basic data exploration which involves finding null values which were zero in the given dataset, and describing the data

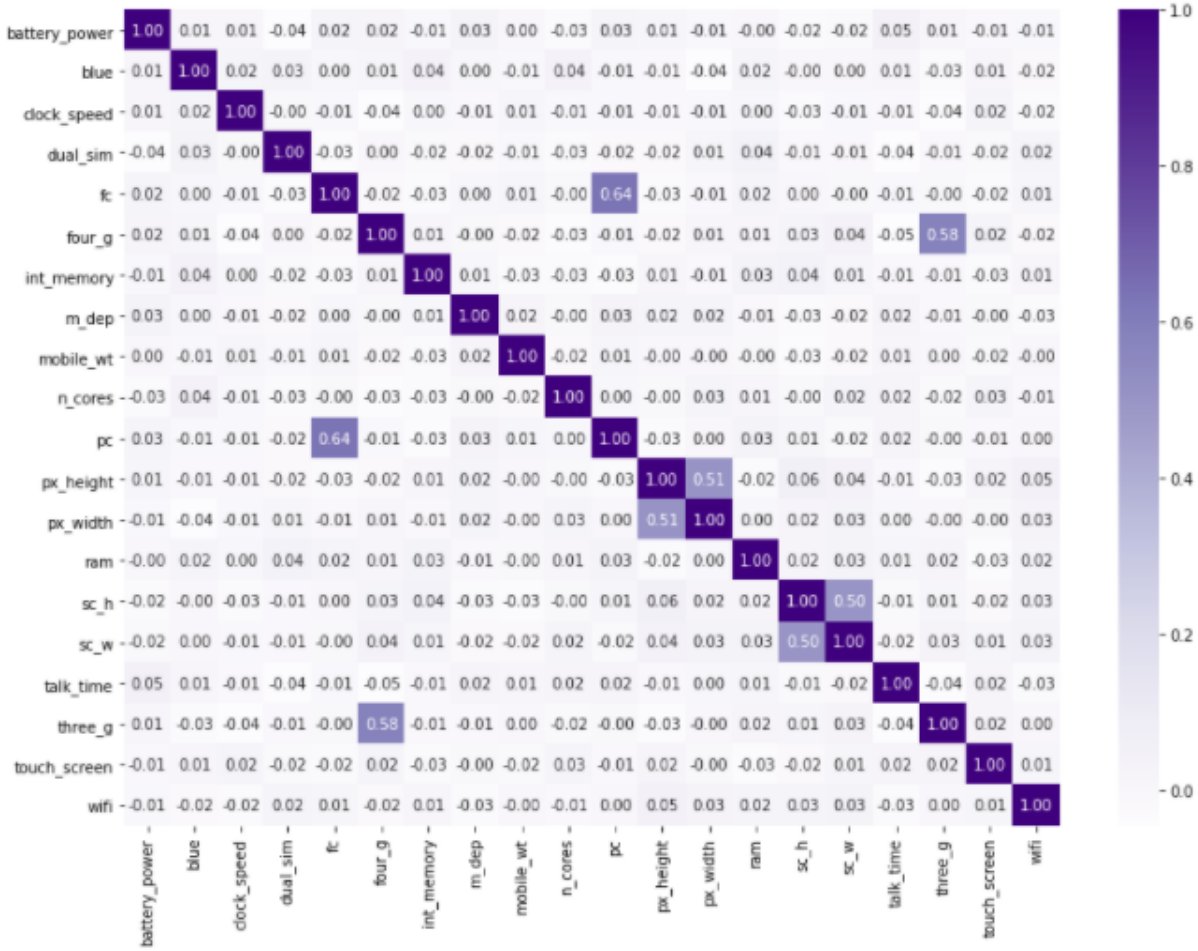
statistically.

	count	mean	std	min	25%	50%	75%	max
battery_power	2000.0	1238.51850	439.418206	501.0	851.75	1226.0	1615.25	1998.0
blue	2000.0	0.49500	0.500100	0.0	0.00	0.0	1.00	1.0
clock_speed	2000.0	1.52225	0.816004	0.5	0.70	1.5	2.20	3.0
dual_sim	2000.0	0.50950	0.500035	0.0	0.00	1.0	1.00	1.0
fc	2000.0	4.30950	4.341444	0.0	1.00	3.0	7.00	19.0
four_g	2000.0	0.52150	0.499662	0.0	0.00	1.0	1.00	1.0
int_memory	2000.0	32.04650	18.145715	2.0	16.00	32.0	48.00	64.0
m_dep	2000.0	0.50175	0.288416	0.1	0.20	0.5	0.80	1.0
mobile_wt	2000.0	140.24900	35.399655	80.0	109.00	141.0	170.00	200.0
n_cores	2000.0	4.52050	2.287837	1.0	3.00	4.0	7.00	8.0
pc	2000.0	9.91650	6.064315	0.0	5.00	10.0	15.00	20.0
px_height	2000.0	645.10800	443.780811	0.0	282.75	564.0	947.25	1960.0
px_width	2000.0	1251.51550	432.199447	500.0	874.75	1247.0	1633.00	1998.0
ram	2000.0	2124.21300	1084.732044	256.0	1207.50	2146.5	3064.50	3998.0
sc_h	2000.0	12.30650	4.213245	5.0	9.00	12.0	16.00	19.0
sc_w	2000.0	5.76700	4.356398	0.0	2.00	5.0	9.00	18.0
talk_time	2000.0	11.01100	5.463955	2.0	6.00	11.0	16.00	20.0
three_g	2000.0	0.76150	0.426273	0.0	1.00	1.0	1.00	1.0
touch_screen	2000.0	0.50300	0.500116	0.0	0.00	1.0	1.00	1.0
wifi	2000.0	0.50700	0.500076	0.0	0.00	1.0	1.00	1.0
price_range	2000.0	1.50000	1.118314	0.0	0.75	1.5	2.25	3.0

We concluded that some of our features were catagorical and some were continuous.

- **Data Visualization:**

  Here our purpose was to find the correlation amongst various features so we created a correlation graph.



  With the graph it is visible that the variables pc and fc, px_width and px_height, three_g and four_g, sc_w and sc_h have high correlations.

- **Feature Engineering:**

  From the correlation matrix we found the variables with high correlations so we created new features that were relevant for our analysis such as screen_size and pixels by taking the square root of sc_w,sc_h and px_height, px_width respectively.

- **Handling Discrepancies:**

  There were certain discrepancies found in the dataset such as in the screen width feature (sc_w) some values were zero which is impractical in real life so to handle zero values, we replaced them with mean of all available values sc_w for all values of sc_h.

- **Outlier Handling:**

  There were two features with outliers and very few values, one of the best ways to handle outliers is choosing a model which can handle outliers. For other models we removed the outliers as there were few in number using quartiles.

- **Data Preprocessing:**

  This step mainly includes scaling the features for the models that require scaling and splitting the dataset to train and test sets for model evaluation and generating the classification report.

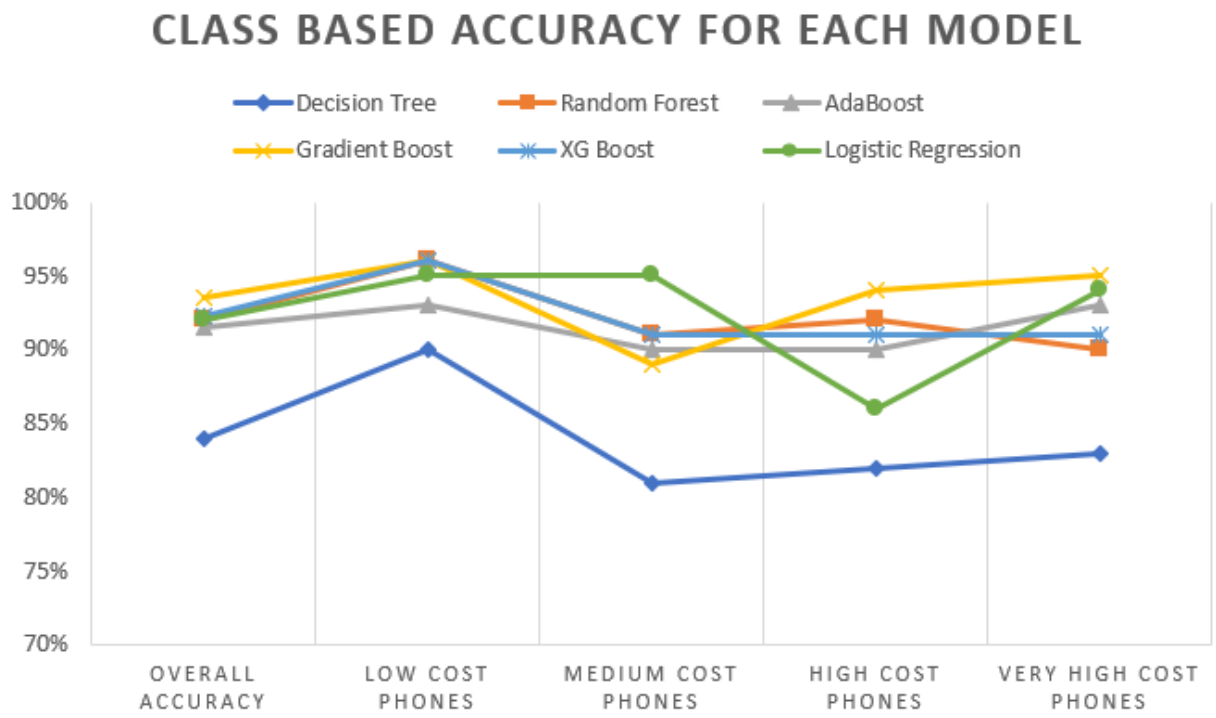
- **Model Selection:**

  We experimented with various models such as SVM, Knn, Logistic Regression, XGBoost, and various others but decided to go with tree based as tree based models have higher interpretability and are not sensitive to outliers.

- **Comparison of accuracies of various models:**

  We compared 6 classifiers and evaluated them based on overall accuracy & class based accuracy as well.

  Decision Trees
  Random Forest
  Ada Boost
  Gradient Boosting
  XGBoost
  Logistic Regression

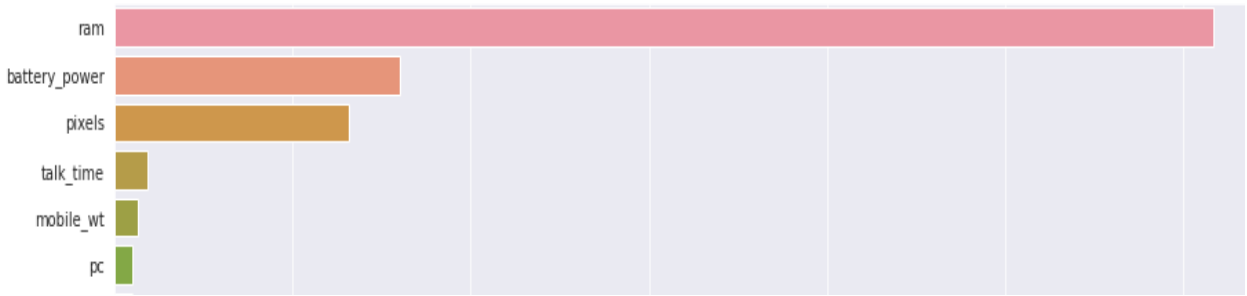


The above graph shows the different accuracies both for overall as well as for each class for various models and we chose to go with XGBoost as it has a good overall accuracy as well as a consistent accuracy for 3 out of the 4 classes.

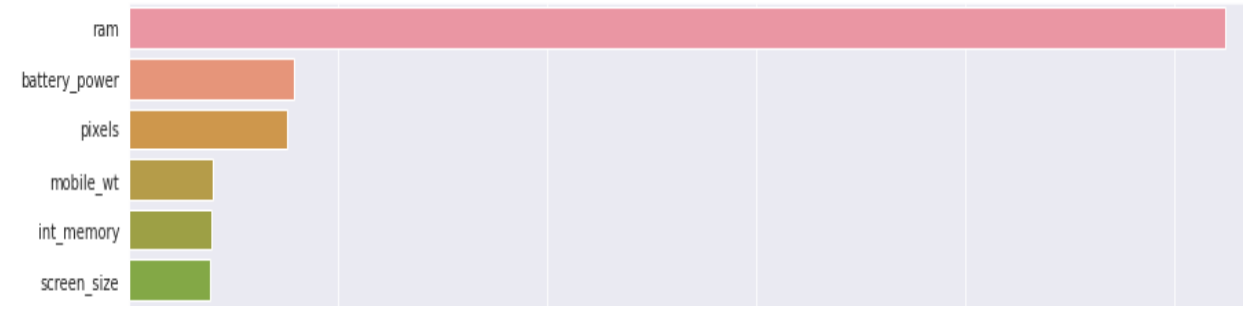
- **Feature Importance:**

Important features from our models are:

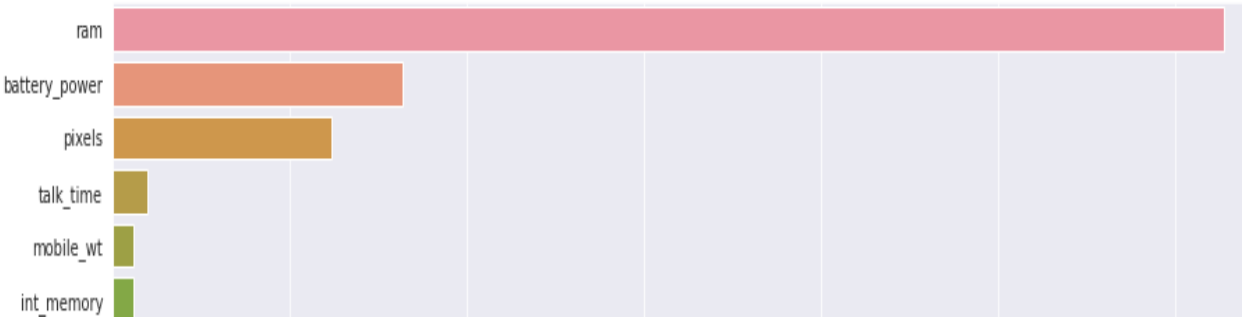
Features from decision tree



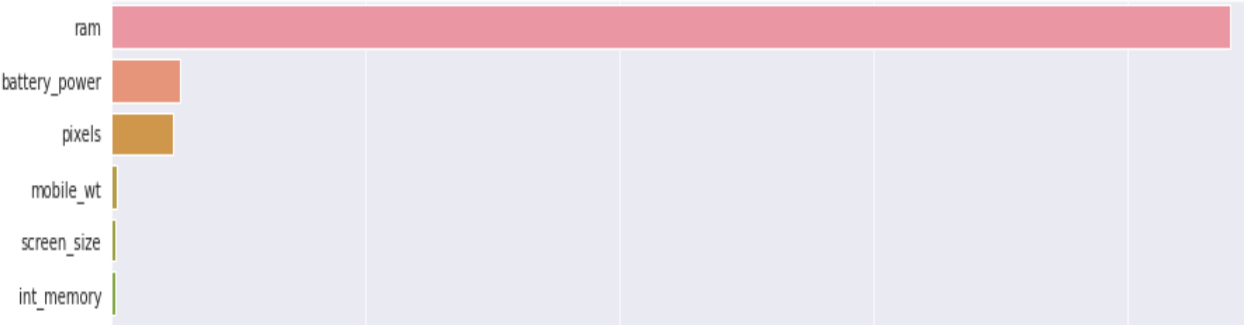
Features from random forest



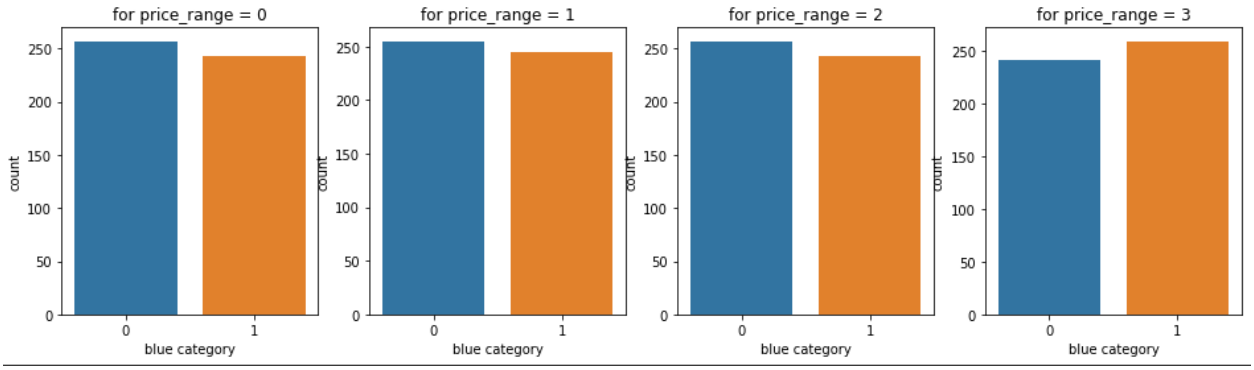
Features from Ada Boost



Features from gradient boost



We got many important features from our models but the most important ones are battery_power and ram.
Apart from selected important features any feature doesn't show variation along the different price ranges. Here is an example of bluetooth.



- **Challenges:**

- We performed "Hypothesis driven EDA" based on domain, but unluckily most of our hypotheses 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.

- **Conclusion:**

- Gradient Boost, Random forest and ADABoost Models are also giving us good overall accuracy but they didn't perform well on Individual classes.
- Out of all the models we have tried XG Boost is performing well on Overall as well as Individual classes.
- Ram, Battery power, Mobile weight, Screen size and pixels are key features in predicting the mobile price range.
- Most of the mis-classifications were encountered between Medium range phones and high range phones. To counter that we can train a specific model for these two classes and can reclassify the cases when the base model predicts the result as Medium range or High range.