Capstone Project-3

Supervised ML- Classification

Topic: Mobile Price Range Predictions

By- Rajat Arya

Contents Of The Presentation:

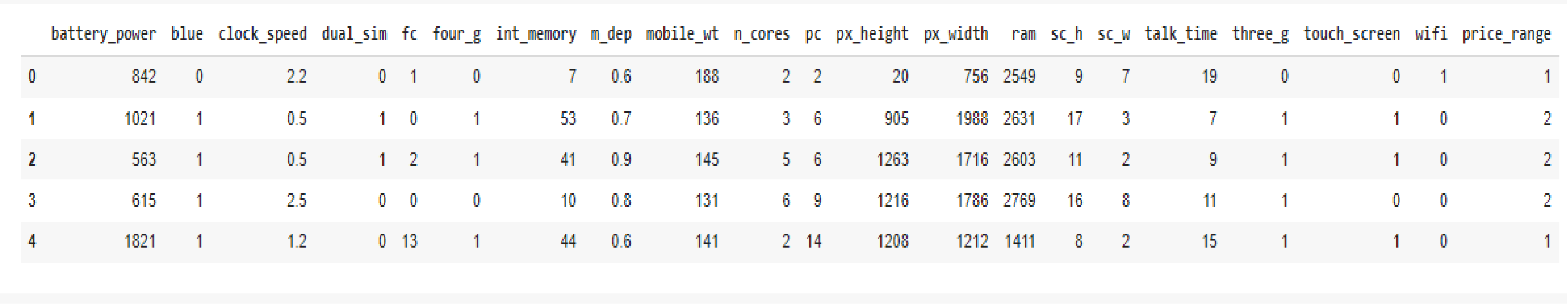
- Defining the problem statement..
- Data Summary
- Data Wrangling
- EDA and Data visualization
- Feature selection
- Applying model
- Model validation and selection
- Challenges
- Conclusion

Problem Statement

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.

Data Summary

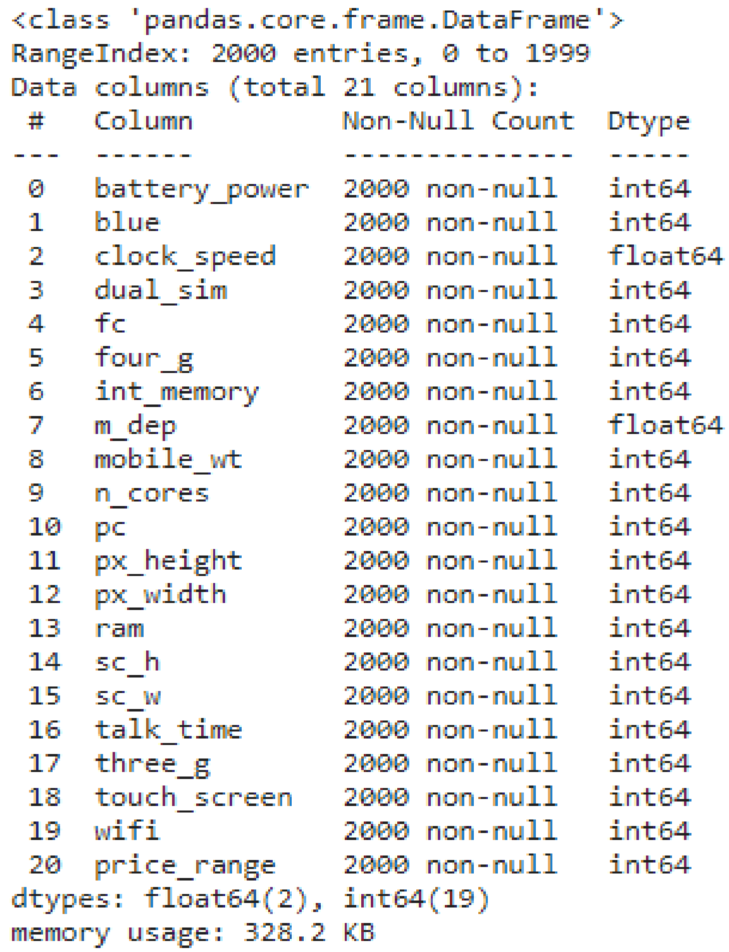
- Size of data: 2000 rows, 21 columns
- Training data: 1600 rows, 12 columns
- Testing data: 400 rows,12 columns



	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
0	842	0	2.2	0	1	0	7	0.6	188	2	2	20	756	2549	9	7	19	0	0	1	1
1	1021	1	0.5	1	0	1	53	0.7	136	3	6	905	1988	2631	17	3	7	1	1	0	2
2	563	1	0.5	1	2	1	41	0.9	145	5	6	1263	1716	2603	11	2	9	1	1	0	2
3	615	1	2.5	0	0	0	10	0.8	131	6	9	1216	1786	2769	16	8	11	1	0	0	2
4	1821	1	1.2	0	13	1	44	0.6	141	2	14	1208	1212	1411	8	2	15	1	1	0	1

Data Summary

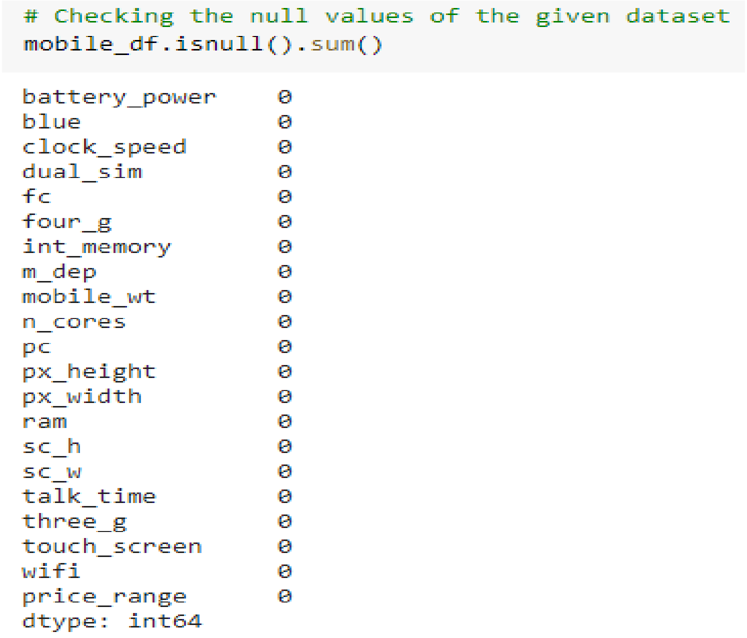
- Battery_power - Total energy a battery can store in one time measured in mAh
- Blue - Has Bluetooth or not
- Clock_speed - speed at which microprocessor executes instructions
- Dual_sim - Has dual sim support or not
- Fc - Front Camera mega pixels
- Four_g - Has 4G or not
- Int_memory - Internal Memory 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 - Pixel Resolution Height
- Px_width - Pixel Resolution Width
- Ram - Random Access Memory in Mega Bytes
- 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 value of 0(low cost), 1(medium cost),2(high cost) and 3(very high cost).



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2000 entries, 0 to 1999
Data columns (total 21 columns):
 #   Column         Non-Null Count  Dtype  
---  ------         --------------  -----  
 0   battery_power  2000 non-null   int64  
 1   blue           2000 non-null   int64  
 2   clock_speed    2000 non-null   float64
 3   dual_sim       2000 non-null   int64  
 4   fc             2000 non-null   int64  
 5   four_g         2000 non-null   int64  
 6   int_memory     2000 non-null   int64  
 7   m_dep          2000 non-null   float64
 8   mobile_wt      2000 non-null   int64  
 9   n_cores        2000 non-null   int64  
 10  pc             2000 non-null   int64  
 11  px_height      2000 non-null   int64  
 12  px_width       2000 non-null   int64  
 13  ram            2000 non-null   int64  
 14  sc_h           2000 non-null   int64  
 15  sc_w           2000 non-null   int64  
 16  talk_time      2000 non-null   int64  
 17  three_g        2000 non-null   int64  
 18  touch_screen   2000 non-null   int64  
 19  wifi           2000 non-null   int64  
 20  price_range    2000 non-null   int64  
dtypes: float64(2), int64(19)
memory usage: 328.2 KB
```

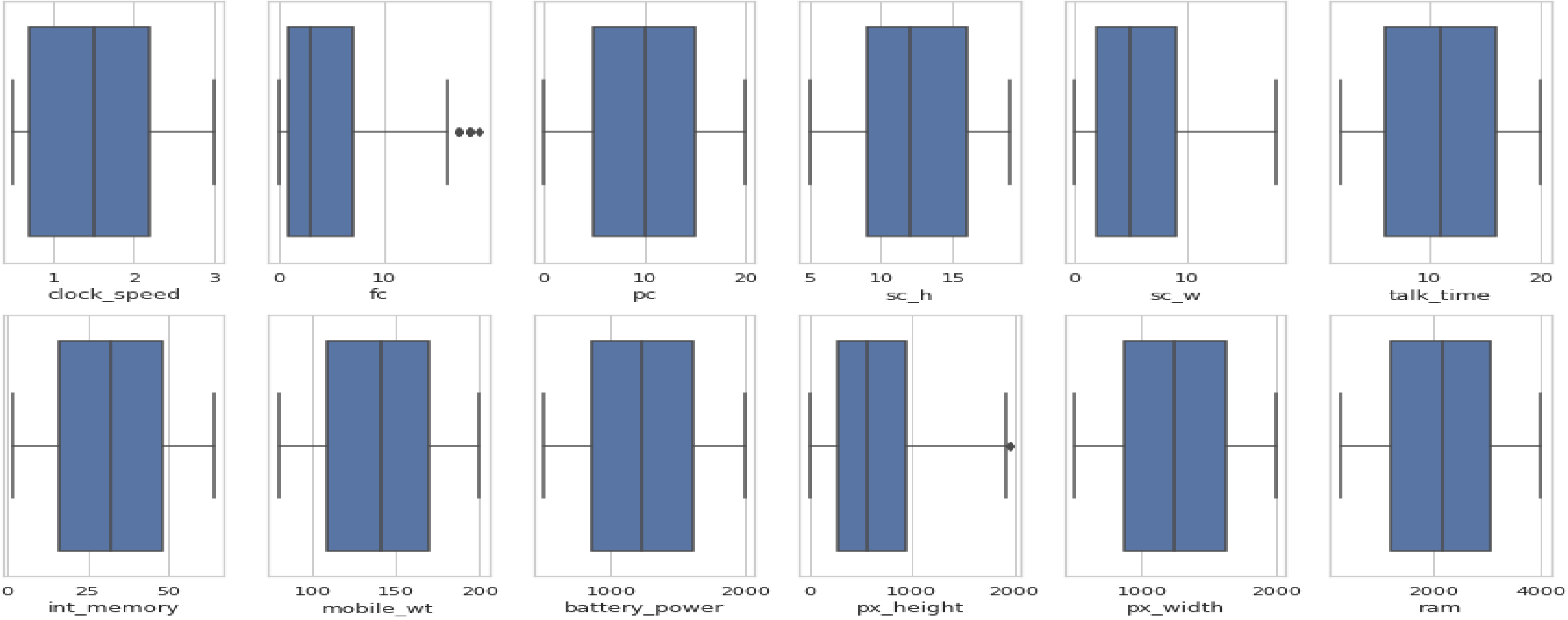
Data Wrangling

- There is no duplicate observation present in our dataset.
- There are no null values in our dataset.
- The boxplot clearly shows that there are no outliers except in fc, which can be considered unimportant because they are not that far away from the maximum value.



```
# Checking the null values of the given dataset
mobile_df.isnull().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
```



Exploratory Data Analysis(EDA) and Data Visualization

- Exploratory Data Analysis (EDA) refers to the critical process of performing initial investigations on a dataset (import from CSV, Jason, and Html files) to summarize their main characteristics, often with visual methods and check assumptions with the help of summary statistics and graphical representations. EDA is used for analyzing what the data can tell us before the modeling or by applying any set of instructions/code. When you are working with the datasets, it is not easy to determine the important characteristics of the data by looking at the column of numbers or a whole spreadsheet/dataset. It may be tedious, boring, and/or overwhelming to determine experiences by seeing plain numbers. Exploratory data analysis techniques have been devised as an aid in this situation.

- Data visualization is the process of translating large data sets and metrics into charts, graphs and other visuals. The resulting visual representation of data makes it easier to identify and share real-time trends, outliers, and new insights about the information represented in the data.

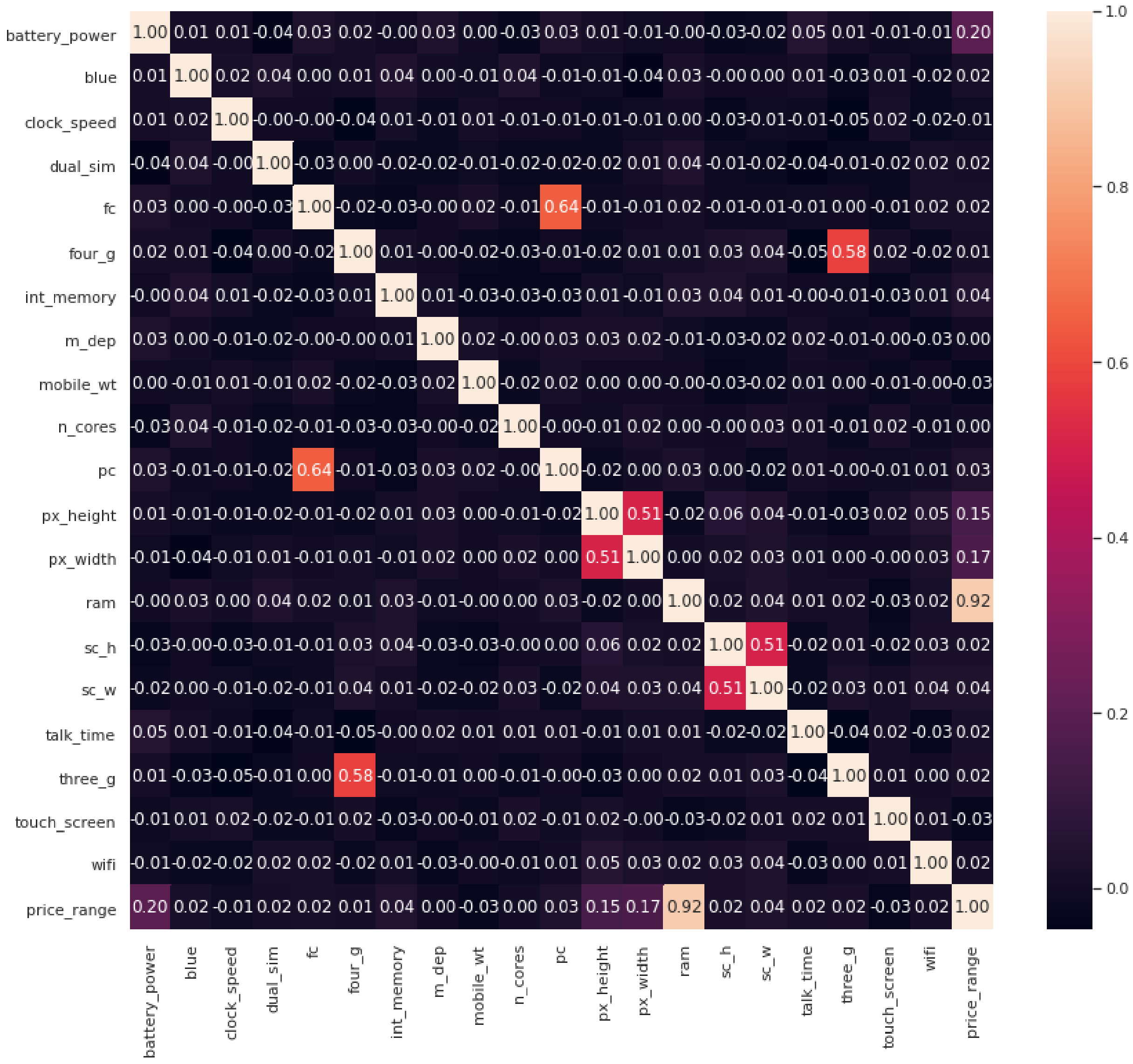
Heat map for independent and dependent variables

- Feature variable Ram is highly correlated with the dependent variable. The variable Ram has more impact on the dependent variable as compare to other variables.

- The independent variable Front Camera and Primary camera are moderately correlated with each other. Similarly, 3G and 4G are also moderately correlated with each other.

- Screen width and screen height are also moderately correlated with each other. Similarly, pixel width and pixel height are moderately correlated with each other.

- We can see that the except ram the other variables battery_power, px_height, px_width also have some impact on the target variable.



	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
battery_power	1.00	0.01	0.01	-0.04	0.03	0.02	-0.00	0.03	0.00	-0.03	0.03	0.01	-0.01	-0.00	-0.03	-0.02	0.05	0.01	-0.01	-0.01	0.20
blue	0.01	1.00	0.02	0.04	0.00	0.01	0.04	0.00	-0.01	0.04	-0.01	-0.01	-0.04	0.03	-0.00	0.00	0.01	-0.03	0.01	-0.02	0.02
clock_speed	0.01	0.02	1.00	-0.00	-0.00	-0.04	0.01	-0.01	0.01	-0.01	-0.01	-0.01	-0.01	0.00	-0.03	-0.01	-0.01	-0.05	0.02	-0.02	-0.01
dual_sim	-0.04	0.04	-0.00	1.00	-0.03	0.00	-0.02	-0.02	-0.01	-0.02	-0.02	-0.02	0.01	0.04	-0.01	-0.02	-0.04	-0.01	-0.02	0.02	0.02
fc	0.03	0.00	-0.00	-0.03	1.00	-0.02	-0.03	-0.00	0.02	-0.01	0.64	-0.01	-0.01	0.02	-0.01	-0.01	-0.01	0.00	-0.01	0.02	0.02
four_g	0.02	0.01	-0.04	0.00	-0.02	1.00	0.01	-0.00	-0.02	-0.03	-0.01	-0.02	0.01	0.01	0.03	0.04	-0.05	0.58	0.02	-0.02	0.01
int_memory	-0.00	0.04	0.01	-0.02	-0.03	0.01	1.00	0.01	-0.03	-0.03	-0.03	0.01	-0.01	0.03	0.04	0.01	-0.00	-0.01	-0.03	0.01	0.04
m_dep	0.03	0.00	-0.01	-0.02	-0.00	-0.00	0.01	1.00	0.02	-0.00	0.03	0.03	0.02	-0.01	-0.03	-0.02	0.02	-0.01	-0.00	-0.03	0.00
mobile_wt	0.00	-0.01	0.01	-0.01	0.02	-0.02	-0.03	0.02	1.00	-0.02	0.02	0.00	0.00	-0.00	-0.03	-0.02	0.01	0.00	-0.01	-0.00	-0.03
n_cores	-0.03	0.04	-0.01	-0.02	-0.01	-0.03	-0.03	-0.00	-0.02	1.00	-0.00	-0.01	0.02	0.00	-0.00	0.03	0.01	-0.01	0.02	-0.01	0.00
pc	0.03	-0.01	-0.01	-0.02	0.64	-0.01	-0.03	0.03	0.02	-0.00	1.00	-0.02	0.00	0.03	0.00	-0.02	0.01	-0.00	-0.01	0.01	0.03
px_height	0.01	-0.01	-0.01	-0.02	-0.01	-0.02	0.01	0.03	0.00	-0.01	-0.02	1.00	0.51	-0.02	0.06	0.04	-0.01	-0.03	0.02	0.05	0.15
px_width	-0.01	-0.04	-0.01	0.01	-0.01	0.01	-0.01	0.02	0.00	0.02	0.00	0.51	1.00	0.00	0.02	0.03	0.01	0.00	-0.00	0.03	0.17
ram	-0.00	0.03	0.00	0.04	0.02	0.01	0.03	-0.01	-0.00	0.00	0.03	-0.02	0.00	1.00	0.02	0.04	0.01	0.02	-0.03	0.02	0.92
sc_h	-0.03	-0.00	-0.03	-0.01	-0.01	0.03	0.04	-0.03	-0.03	-0.00	0.00	0.06	0.02	0.02	1.00	0.51	-0.02	0.01	-0.02	0.03	0.02
sc_w	-0.02	0.00	-0.01	-0.02	-0.01	0.04	0.01	-0.02	-0.02	0.03	-0.02	0.04	0.03	0.04	0.51	1.00	-0.02	0.03	0.01	0.04	0.04
talk_time	0.05	0.01	-0.01	-0.04	-0.01	-0.05	-0.00	0.02	0.01	0.01	0.01	-0.01	0.01	0.01	-0.02	-0.02	1.00	-0.04	0.02	-0.03	0.02
three_g	0.01	-0.03	-0.05	-0.01	0.00	0.58	-0.01	-0.01	0.00	-0.01	-0.00	-0.03	0.00	0.02	0.01	0.03	-0.04	1.00	0.01	0.00	0.02
touch_screen	-0.01	0.01	0.02	-0.02	-0.01	0.02	-0.03	-0.00	-0.01	0.02	-0.01	0.02	-0.00	-0.03	-0.02	0.01	0.02	0.01	1.00	0.01	-0.03
wifi	-0.01	-0.02	-0.02	0.02	0.02	-0.02	0.01	-0.03	-0.00	-0.01	0.01	0.05	0.03	0.02	0.03	0.04	-0.03	0.00	0.01	1.00	0.02
price_range	0.20	0.02	-0.01	0.02	0.02	0.01	0.04	0.00	-0.03	0.00	0.03	0.15	0.17	0.92	0.02	0.04	0.02	0.02	-0.03	0.02	1.00

Relation Between Price Range & Ram

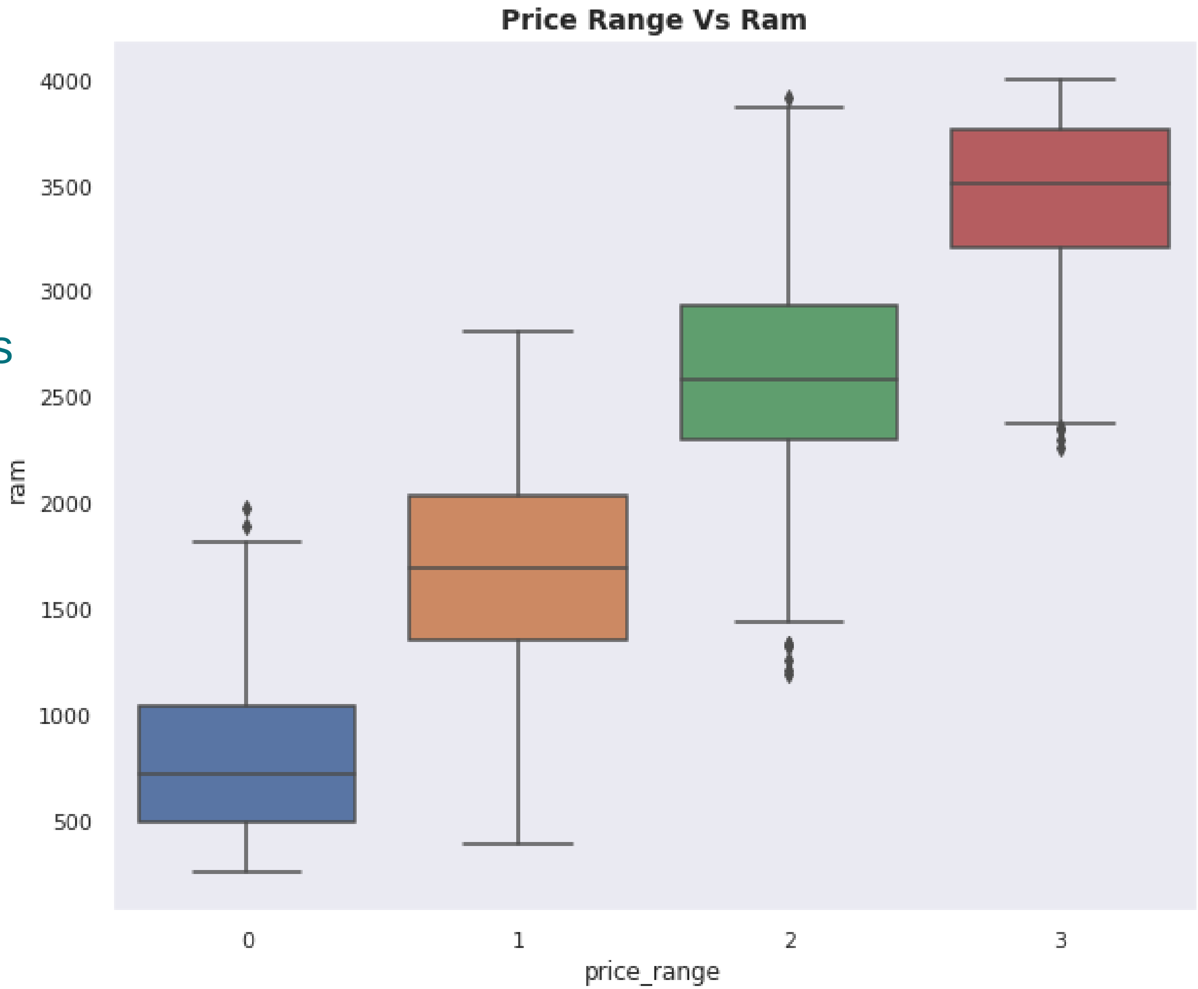
- This is a positive relationship, with increase in RAM, price too increases. There are 4 types of price range

- Type 1(low cost): RAM ranges between 216 to 1974 megabytes

- Type 2(medium cost): RAM ranges between 387 to 2811 megabytes

- Type 3(high cost): RAM ranges between 1185 to 3916 megabytes

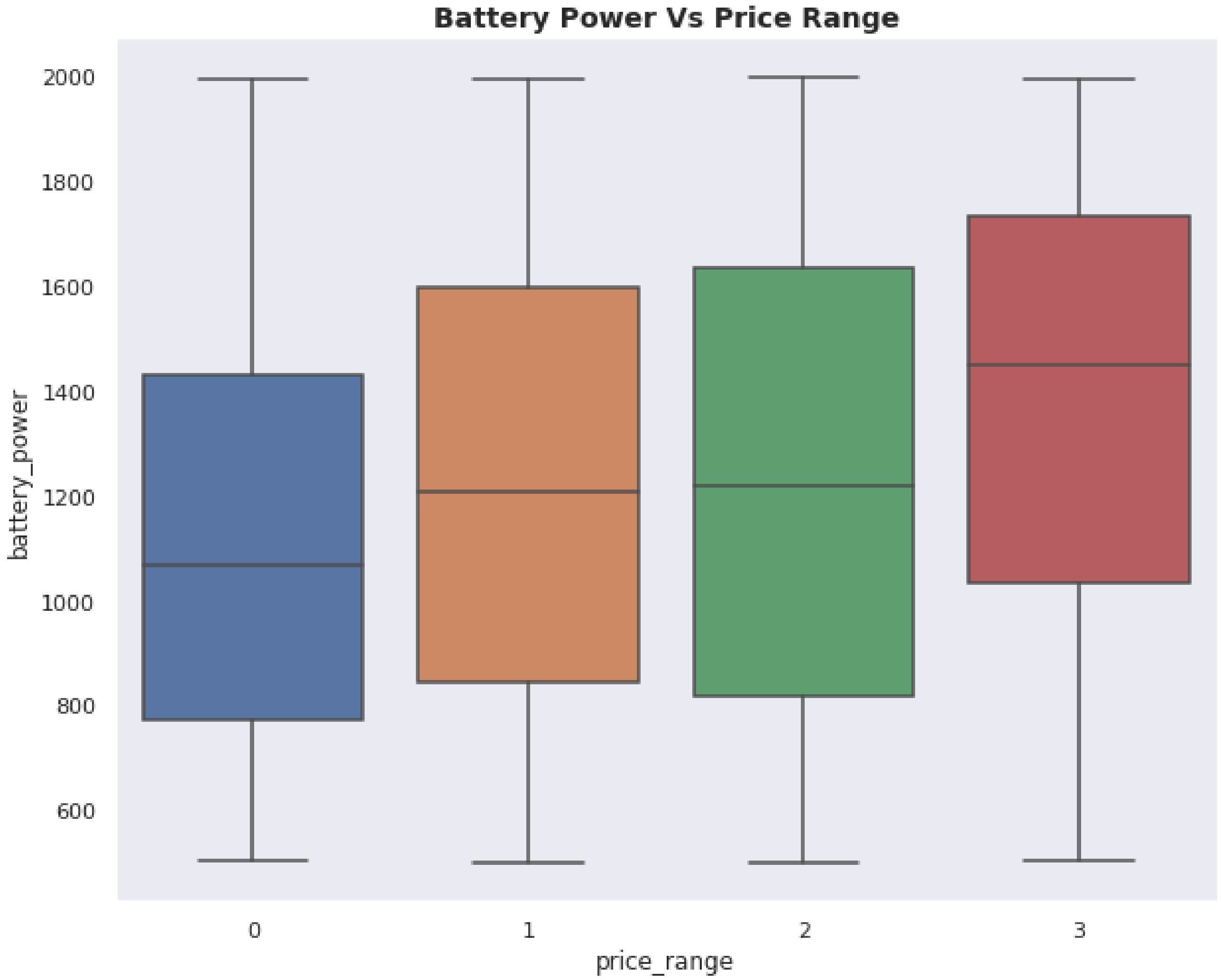
- Type 4(very high cost): RAM ranges between 2255 to 4000 megabytes



Relation between Price Range and Battery Power

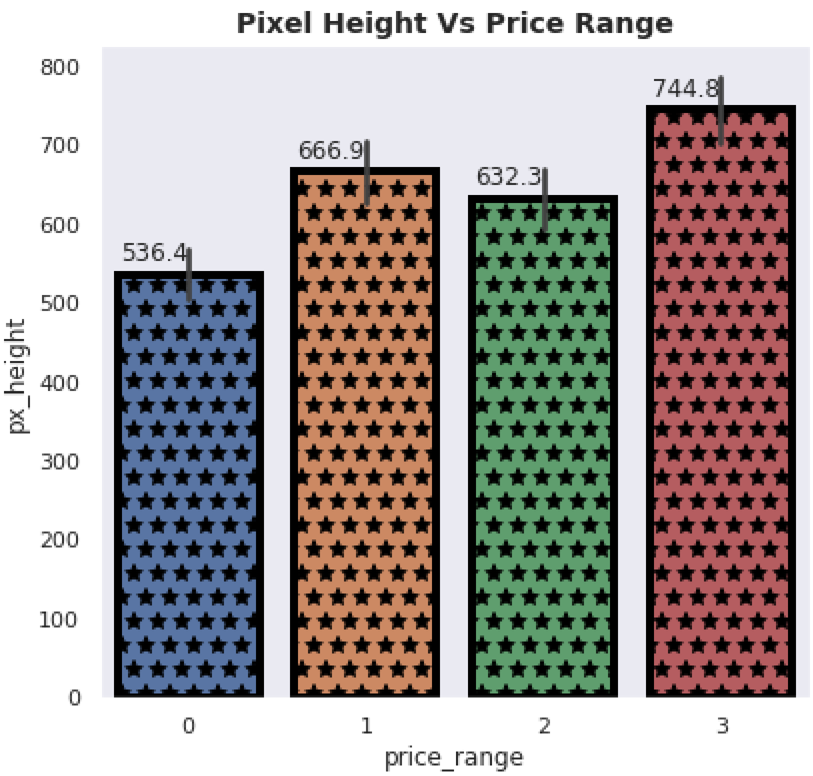
- We can observe from this Boxplot of battery power vs. price range that price range gradually increases as battery power increases. As a result, we can claim that battery power has a positive influence on prediction

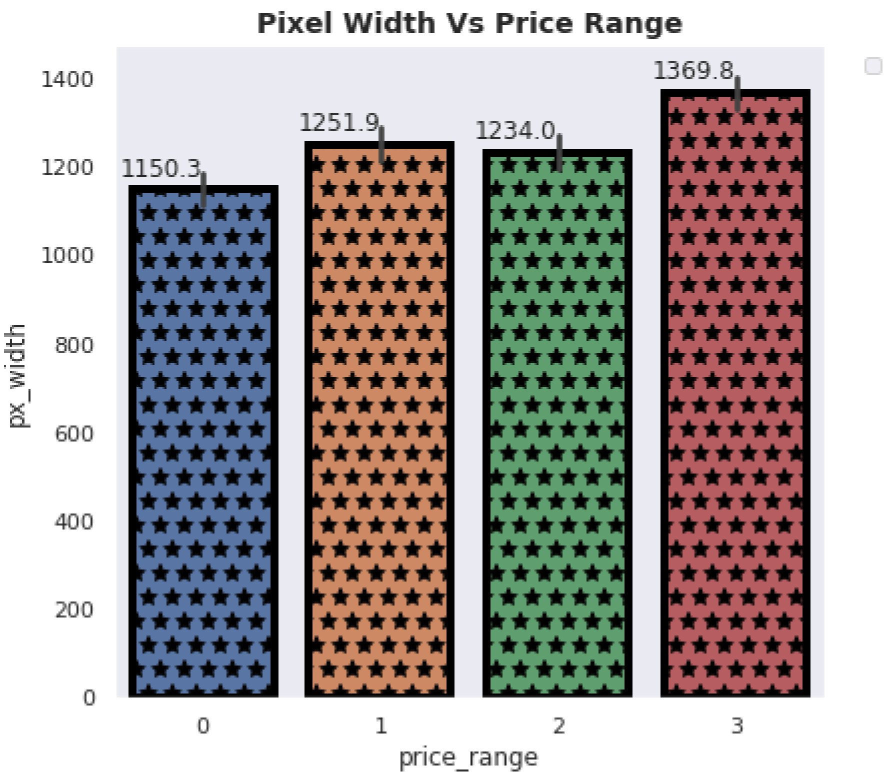
- Average battery power is near about 1200 mah.



Relationship between the Price Range and Pixel Height/ Width

- From the above bar plot, we can see that the average pixel height and width are highest for the price range 3(very high cost).
- Low-cost phones have smaller average pixel width and pixel height.
- We can observe from this Bar plot that pixel height and pixel width are roughly equal in relevance when it comes to model development for prediction.

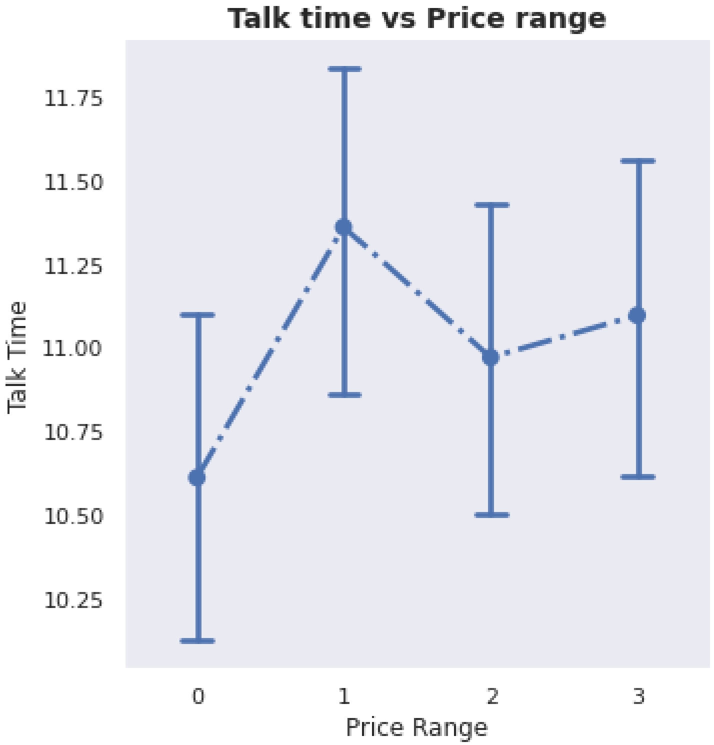

Pixel Height Vs Price Range

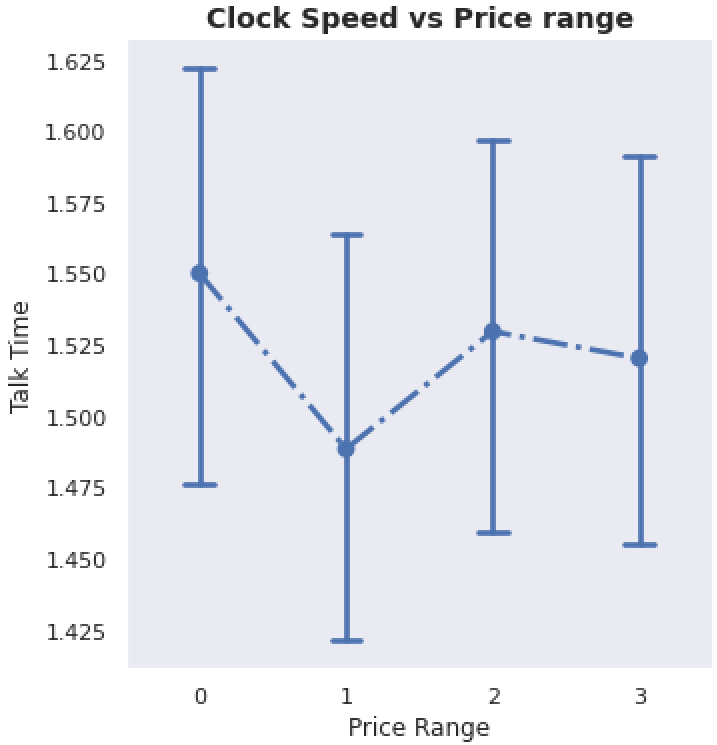

Pixel Width Vs Price Range

Talk Time and Clock Speed Relation with Price Range

- We can see from this Point plot of chat time vs. price range that the price range does not grow steadily as the talk time increases. As a result, we can conclude that talk time has little impact on prediction.

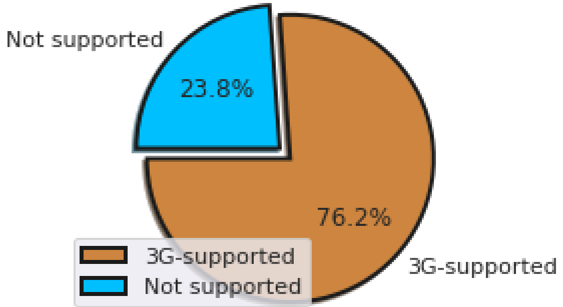
- From this Point plot of clock speed versus price range, we can observe that price range does not progressively increase as clock speed increases. As a result, we can conclude that clock speed had little impact on the prediction.

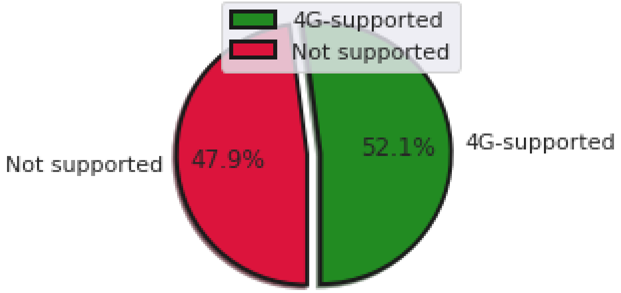


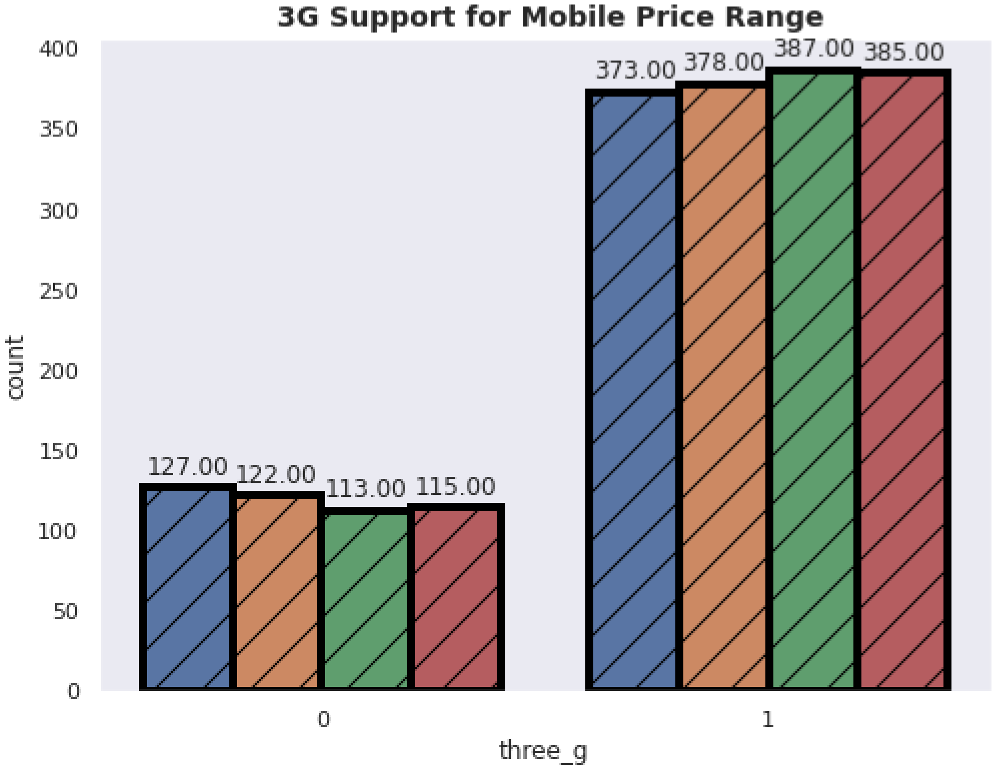


3G And 4G Relationship with Price

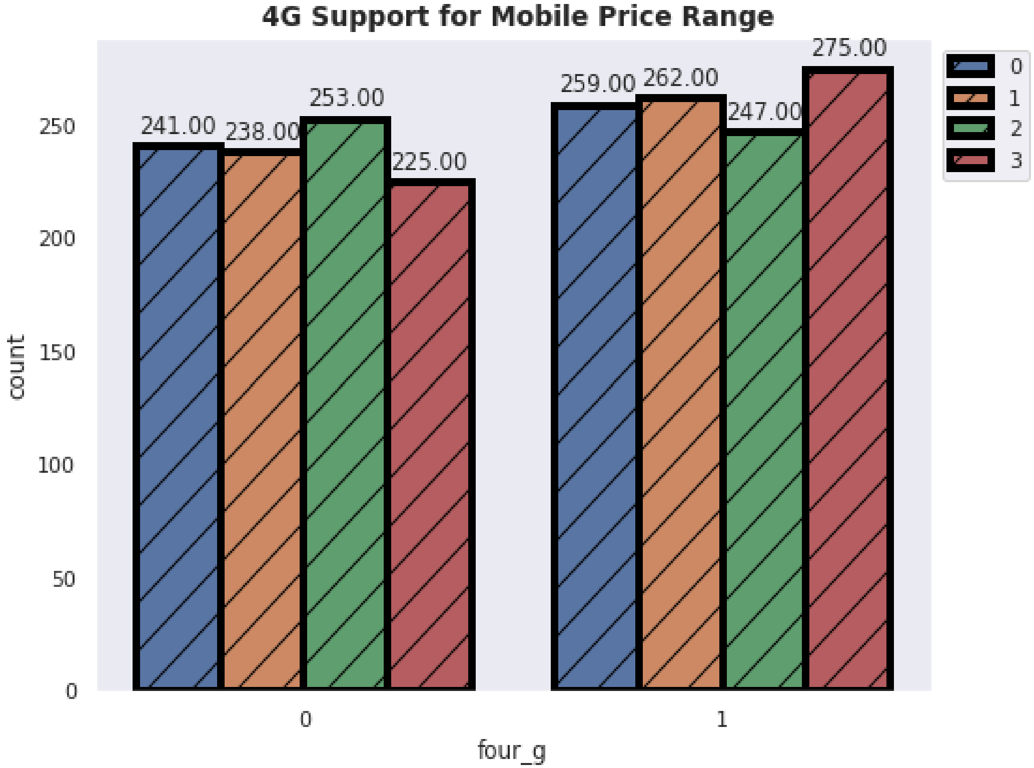
We can observe from the plot that variable three_g has little influence on price range.







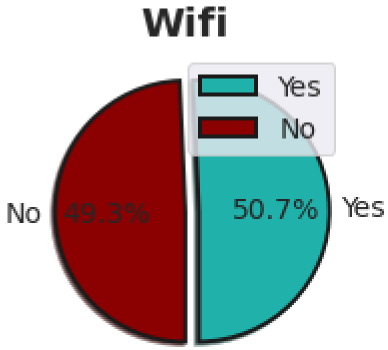
3G Support for Mobile Price Range

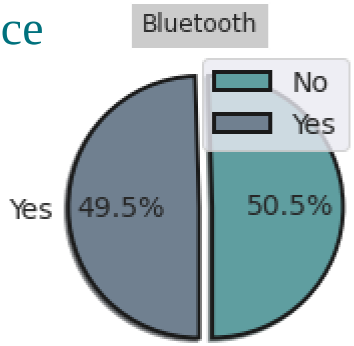


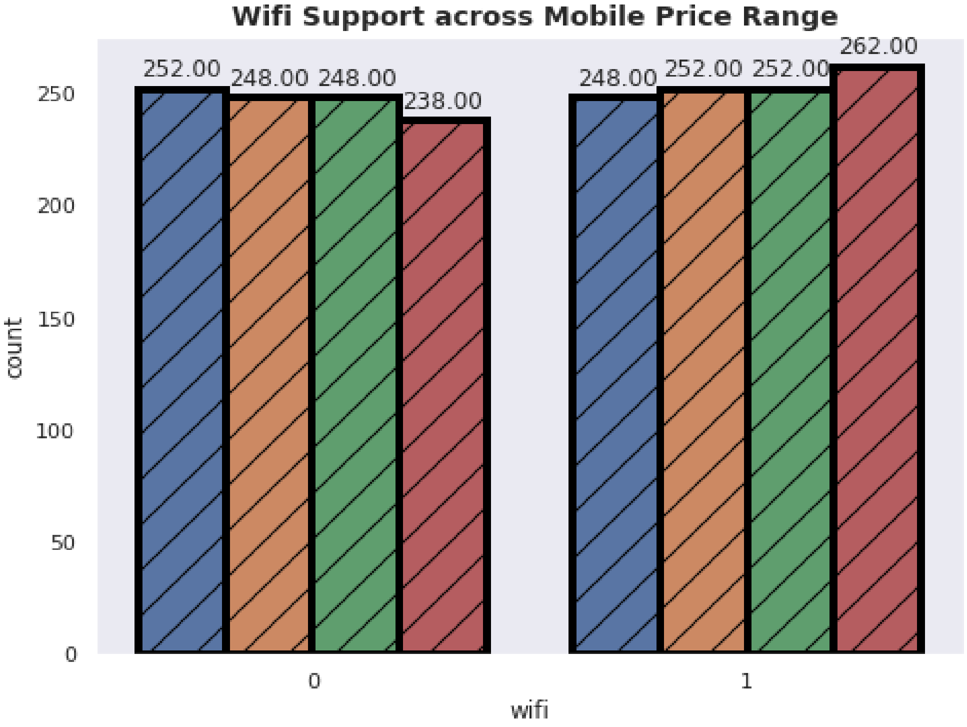
4G Support for Mobile Price Range

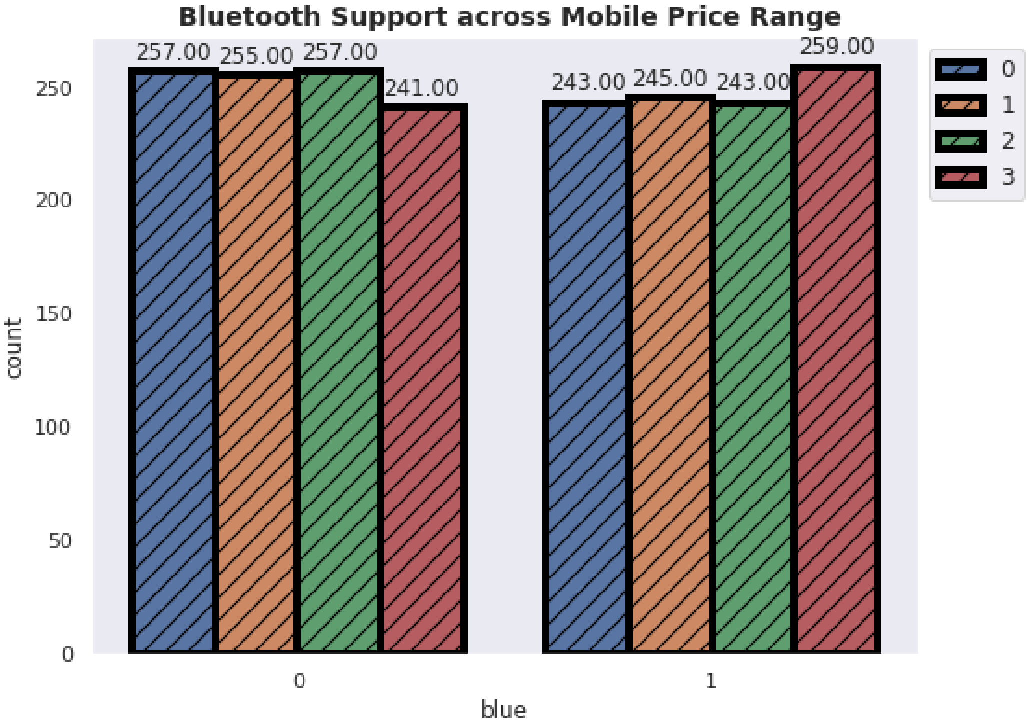
Wi-Fi, And Bluetooth Relationship with Price

We can observe that Wi-Fi and Bluetooth are almost evenly dispersed in both the yes and no categories. As a result, we cannot estimate the price based on these specifications.

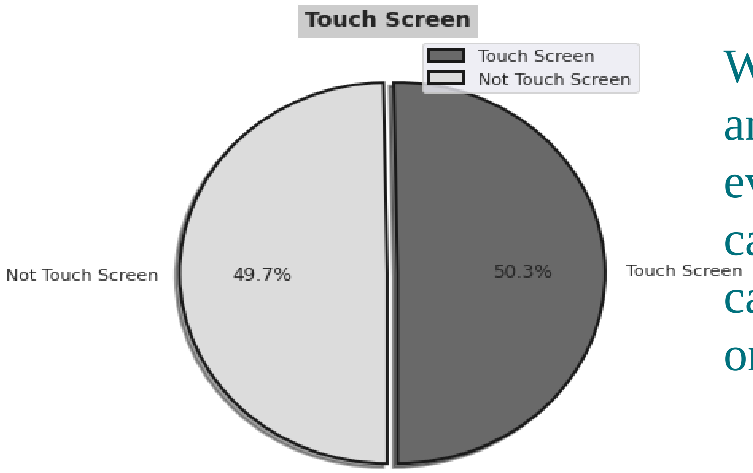




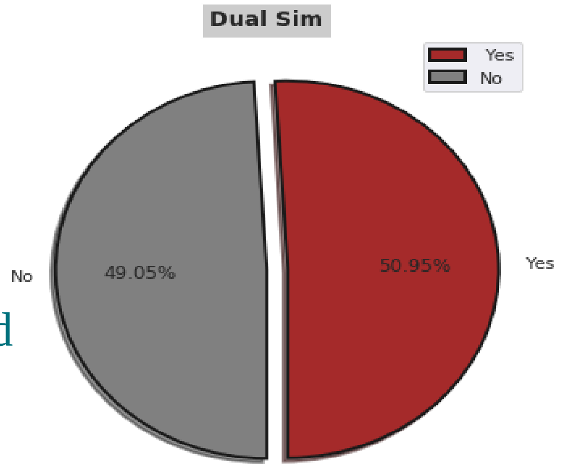


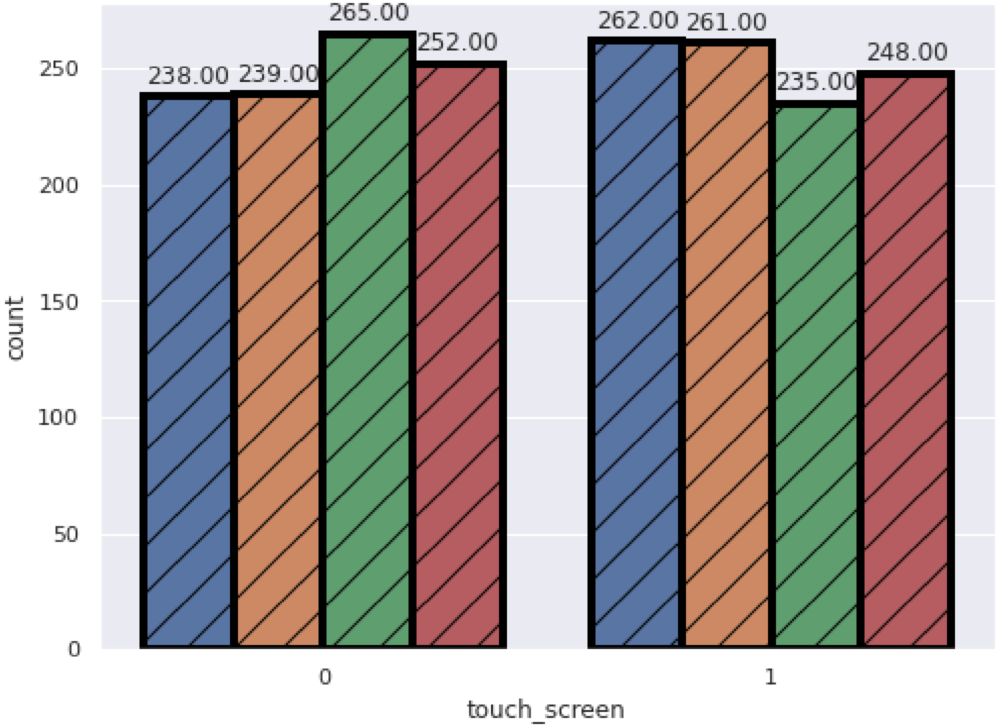


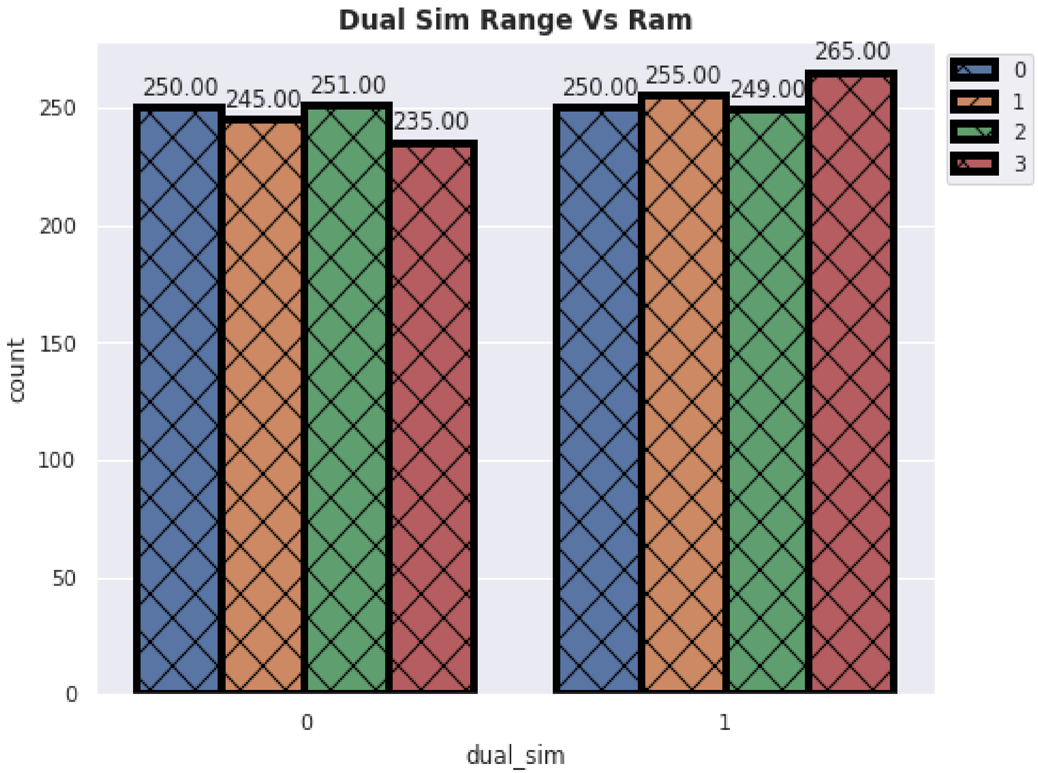
Touch Screen And Dual Sim Relationship With Price



We can observe that Dual Sim and touchscreen are almost evenly dispersed in both the categories. As a result, we cannot estimate the price based on these specifications.

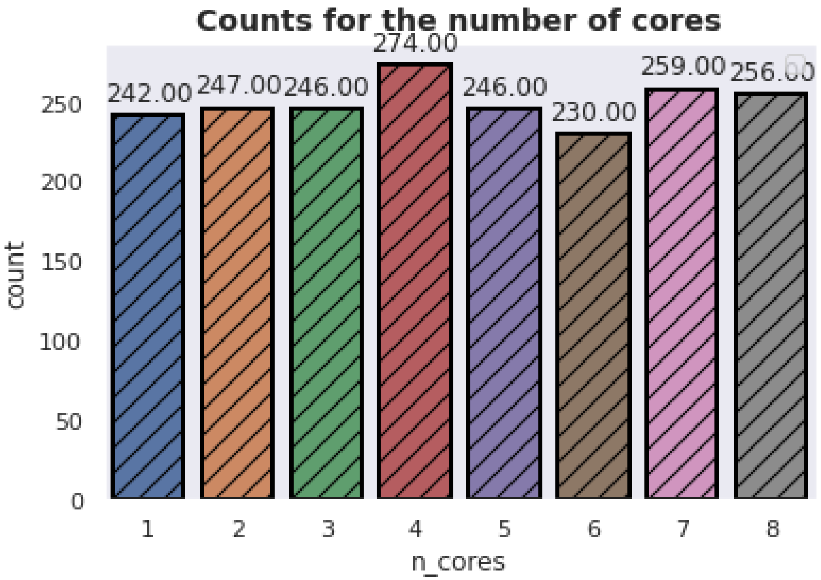


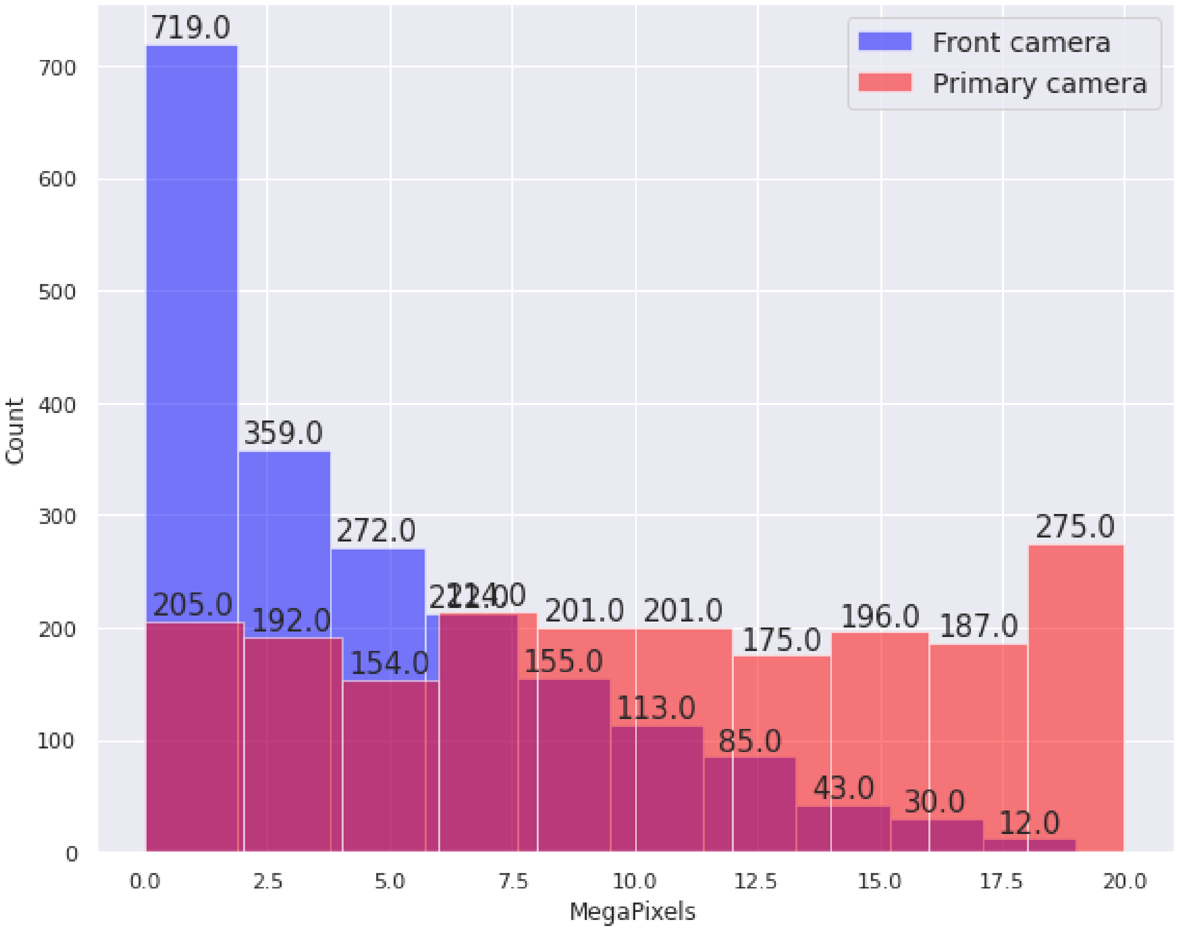




Front Camera Primary Camera and Number of Cores

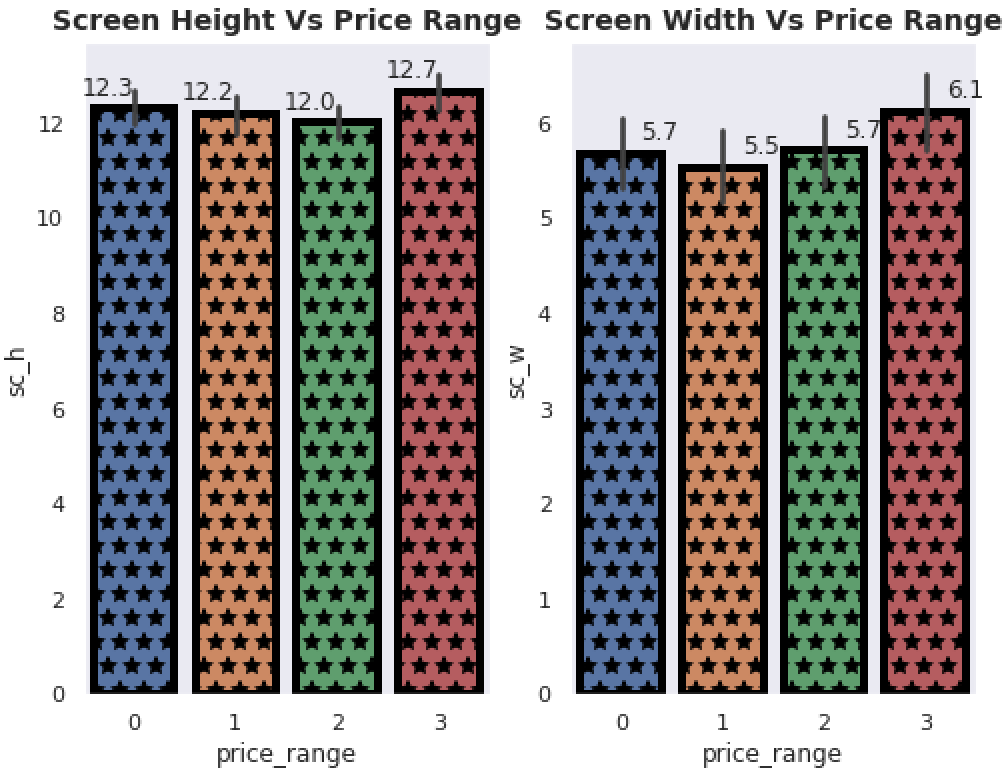
- From the graph, we can see that the Front Camera has more impact on a price range as compared to the primary camera.
- According to the above count plot, the quad core(4 cores) category has the most mobiles.





Plotting the relationship between Screen width/ height and price range

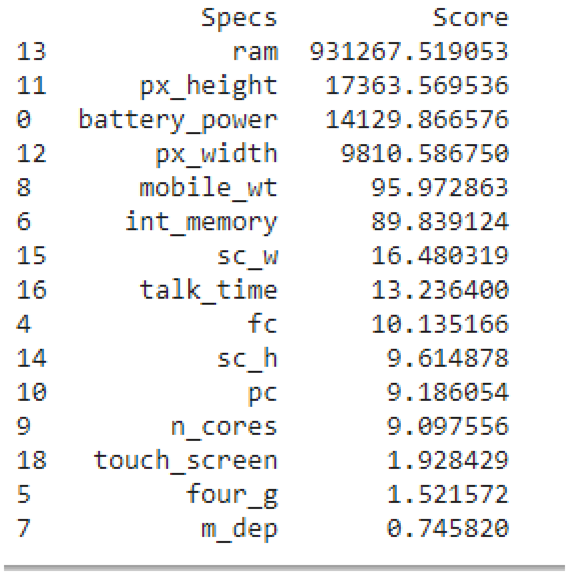
- From the above bar plot, we can see that the average screen height and width are highest for the very high-cost price range.
- High-cost phones have smaller average Screen height.



Feature Selection

In this case, we utilize the SelectKbest method to compute the scores of the top 15 independent variables in relation to the dependent variables.

As a result, we only consider features with scores greater than nine.



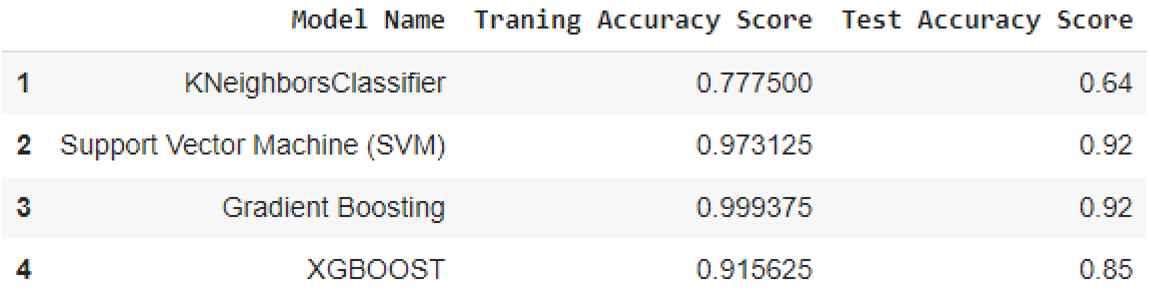
```
             Specs          Score
13             ram  931267.519053
11       px_height   17363.569536
0    battery_power   14129.866576
12        px_width    9810.586750
8        mobile_wt      95.972863
6       int_memory      89.839124
15            sc_w      16.480319
16       talk_time      13.236400
4               fc      10.135166
14            sc_h       9.614878
10              pc       9.186054
9          n_cores       9.097556
18    touch_screen       1.928429
5           four_g       1.521572
7            m_dep       0.745820
```

Applying the Models

- K-Nearest Neighbors(KNN)
- Support Vector Machine (SVM)
- Gradient Boosting
- XGBOOST

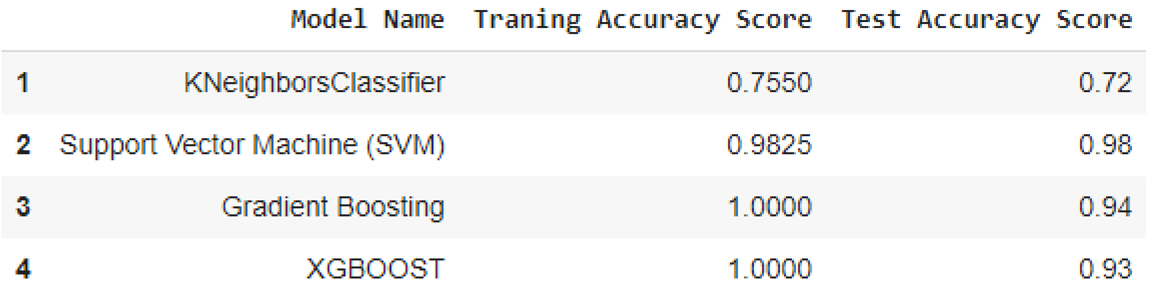
Model Selection And Validation

- Accuracy before the Hyperparameter tuning.



	Model Name	Traning Accuracy Score	Test Accuracy Score
1	KNeighborsClassifier	0.777500	0.64
2	Support Vector Machine (SVM)	0.973125	0.92
3	Gradient Boosting	0.999375	0.92
4	XGBOOST	0.915625	0.85

- Accuracy after Hyperparameter tuning.



	Model Name	Traning Accuracy Score	Test Accuracy Score
1	KNeighborsClassifier	0.7550	0.72
2	Support Vector Machine (SVM)	0.9825	0.98
3	Gradient Boosting	1.0000	0.94
4	XGBOOST	1.0000	0.93

- We chose the Support Vector Machine because of its great accuracy. The best hyperparameters for the SVM are:

The best hyperparameter for Support Vector Machine : {'C': 1000, 'gamma': 0.001, 'kernel': 'rbf'}

Challenges

- Finding an appropriate collection of parameters that could provide us the best results was the most difficult part of this project.

- To determine the best optimal parameters, we must experiment with numerous parameter combinations. It is a time-consuming procedure.

- The correlation matrix shows that the majority of the variables are not associated with the dependent variable. As a result, determining which variable is essential or not is a difficult procedure

.

Conclusion:

- First, we run Data Wrangling on our model to ensure that there are no null or duplicate entries in our dataset. Because there are so few outliers in the independent variable front camera, we can ignore it because it adds no complexity to our model.
- Following the Data Wrangling, we undertake the Exploratory Data Analysis, in which we find the correlation matrix for the dependent and independent variables. According to this correlation matrix, the most essential attributes in terms of mobile pricing range forecasts are Ram, Pixel height, Battery Power, Pixel width, Mobile Weight, and Internal Memory.
- In data visualization, we found that the ram is highly positively correlated with the dependent variable. If ram size increases then the mobile price also increases.
- In the next step, we perform the feature selection using the SelectKbest and take only those features whose score is greater than 9.
- For the prediction, we employ the four models. We chose the support vector machine over all other models since SVM test accuracy is 98 percent. Even after adjusting, the gradient boosting model may be overfitting because there is no discernible difference in test accuracy when using alternative sets of hyperparameters. With hyperparameter adjustment, overfitting in XGBoost is reduced, although accuracy is lower than in SVM.