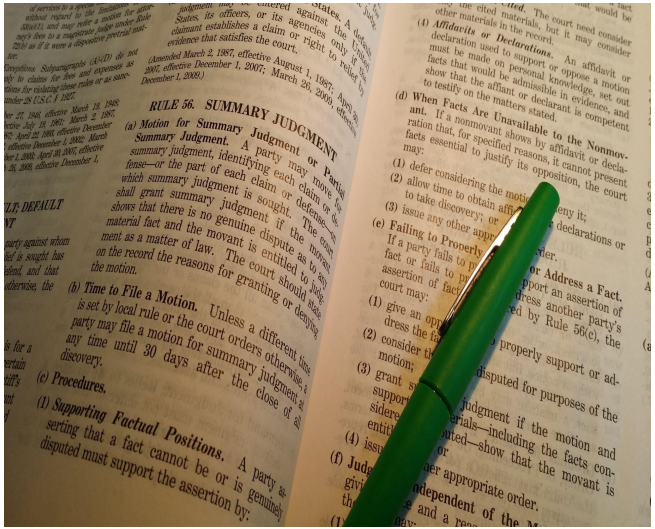
Capstone Project 3
Mobile Price Range Prediction

Team Members

Radhika R Menon
Jayprakash Kunduru
Ghanal Kaushik

Predicting the mobile price range

1. Problem Statement
2. Data Overview and preprocessing
3. Exploratory Data Analysis
4. Feature Engineering
5. Model Implementation
6. Hyperparameter Tuning
7. Model Evaluation Metrics
8. 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. We do not have to predict the actual price but a price range indicating how high the price is i.e. 0 (low cost), 1 (medium cost), 2 (high cost), 3 (very high cost)

Data Overview

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

Data Overview cont.

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

Data Overview cont.

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

Three_g: Has 3g or not

Data Overview contd.

Touch_screen: Has touchscreen 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).

Preprocessing

On Observing the data, it was found out that:
Px_height (pixel height) and sc_w (screen width) have some values equal to 0.

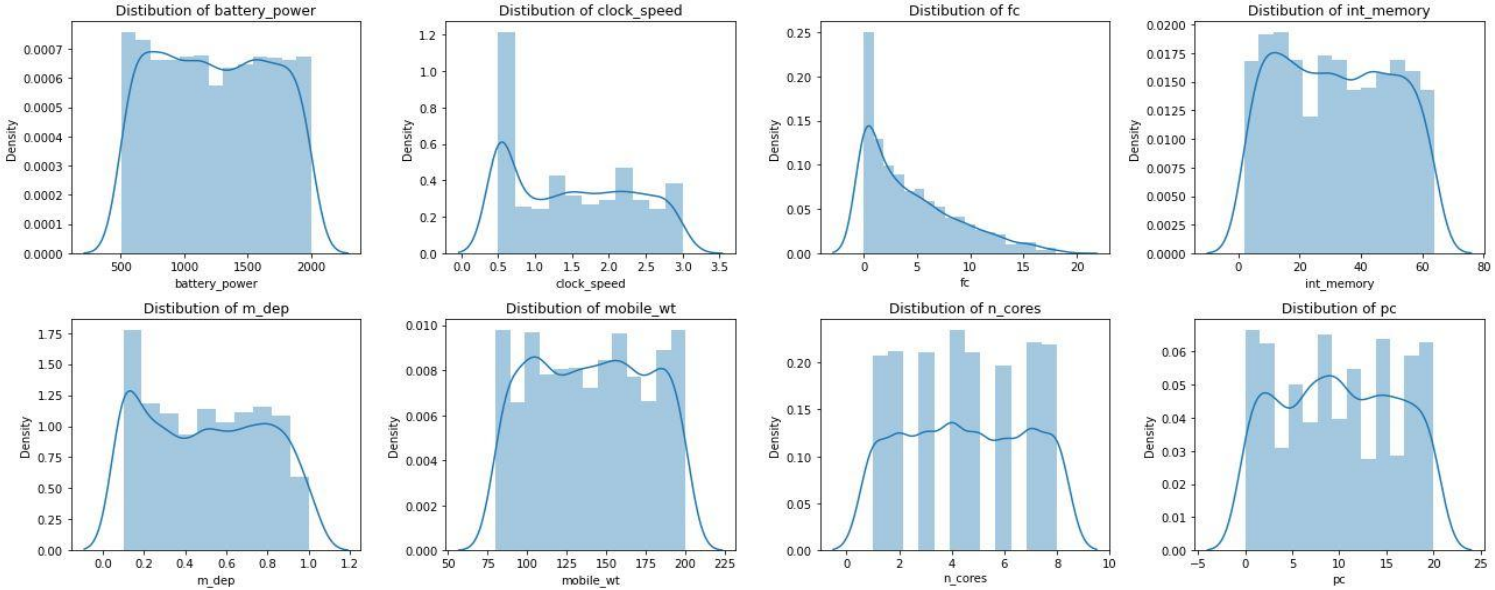
So replaced those values with the median values of their respective columns.

	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

EDA

Univariate Analysis

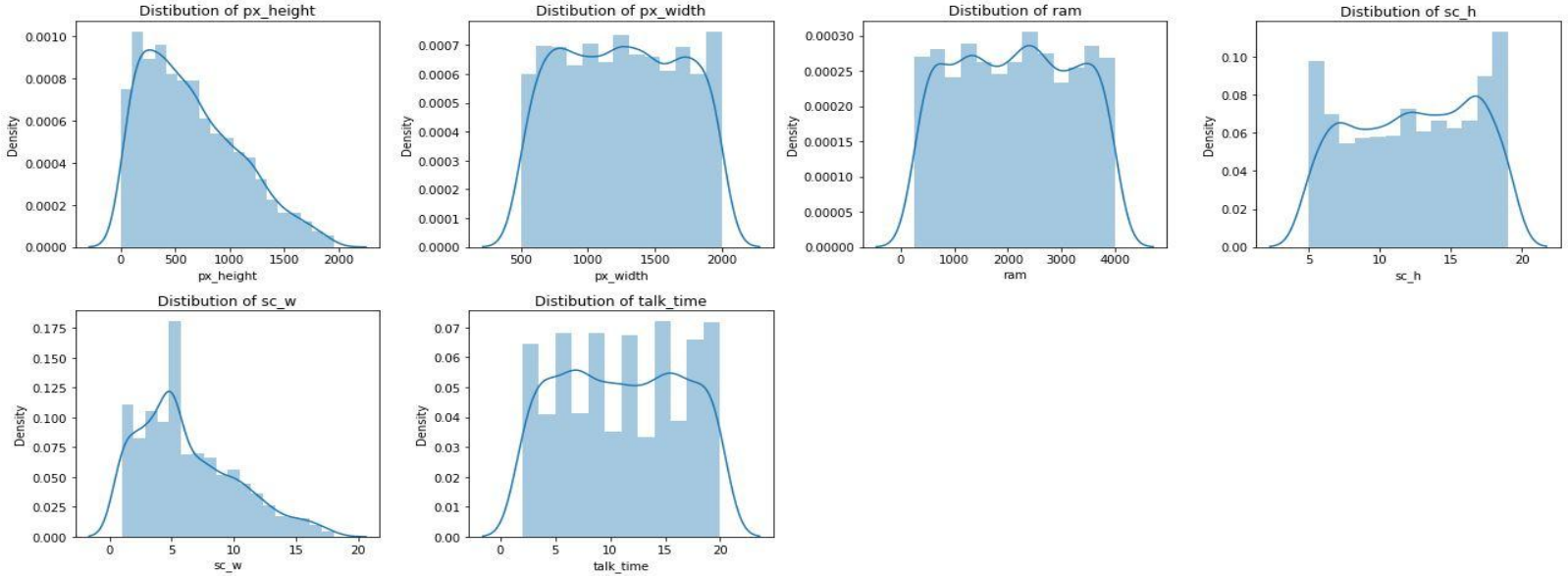
Distribution of features:



EDA

Univariate Analysis contd.

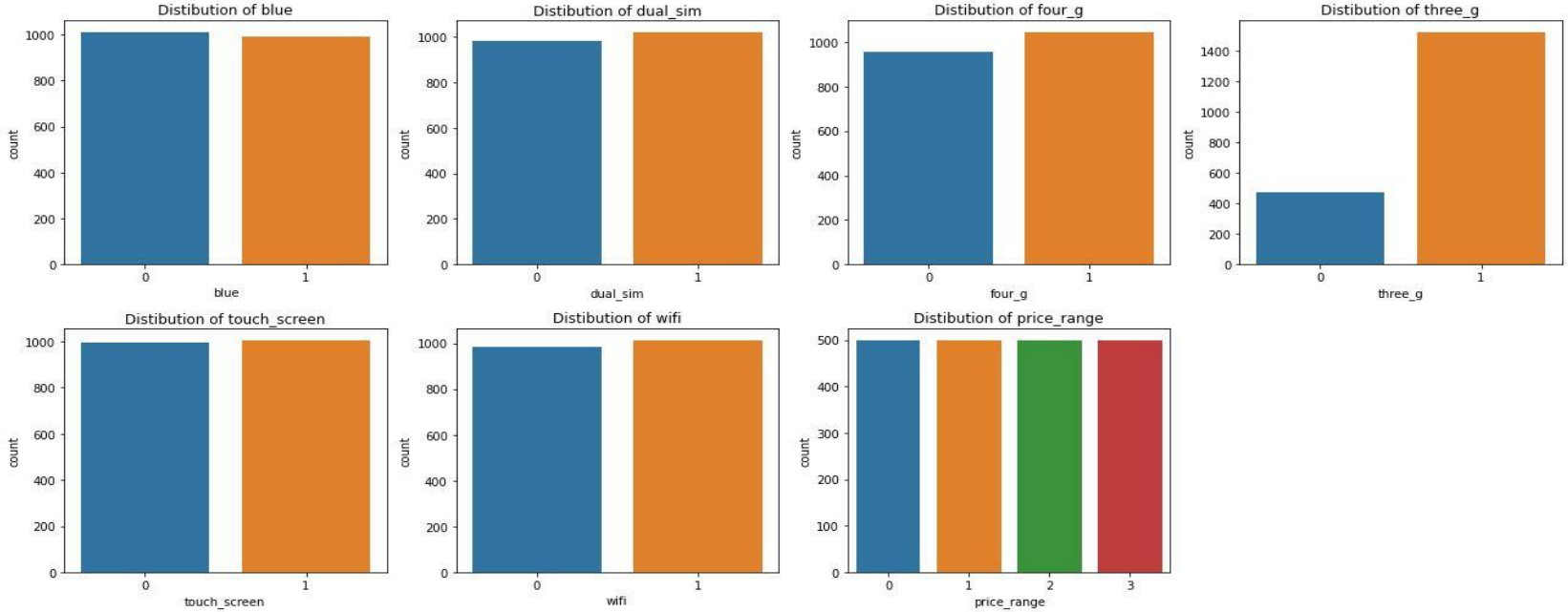
Distribution of features:



EDA

Univariate Analysis contd.

Distribution of features:



EDA

Univariate Analysis contd.

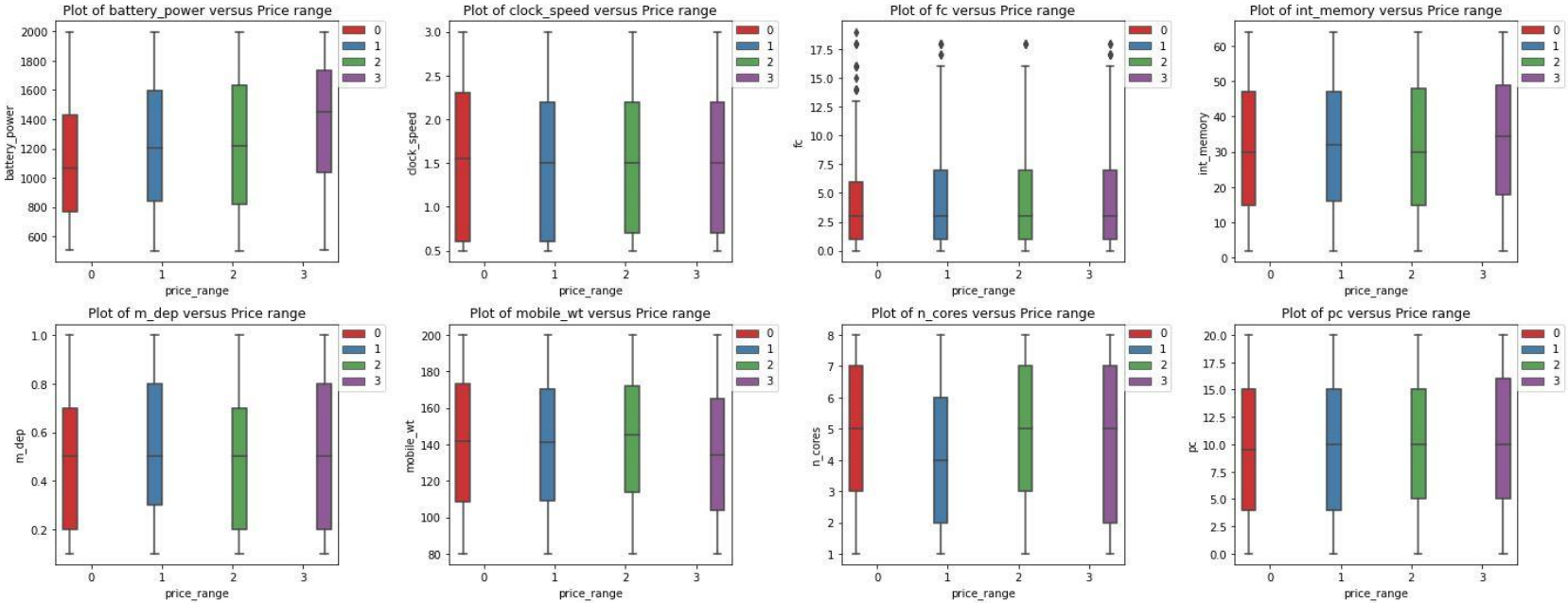
Conclusion:

- **Most of the phones currently available in the market have 3G support.**
- **There is a similar count of records for phones belonging to all price ranges in the given data.**
- **Most of the don't contain front camera or having low quality cameras.**

EDA

Bivariate Analysis

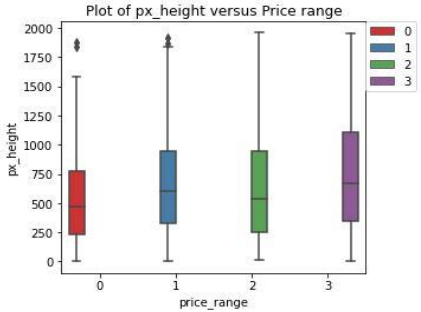
Effect of Target variable on features:

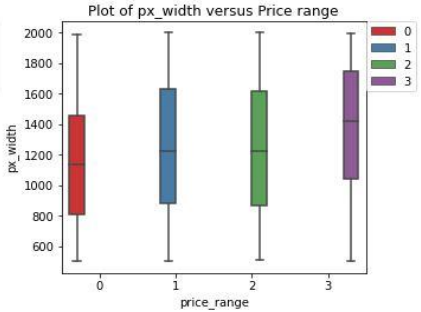


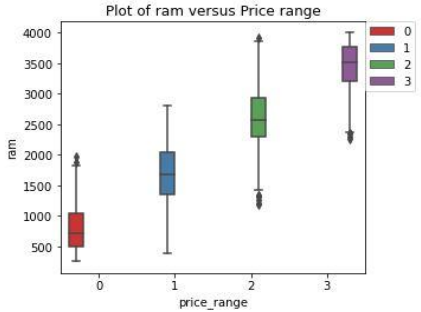
EDA

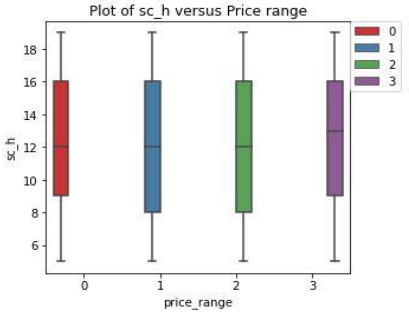
Bivariate Analysis contd.

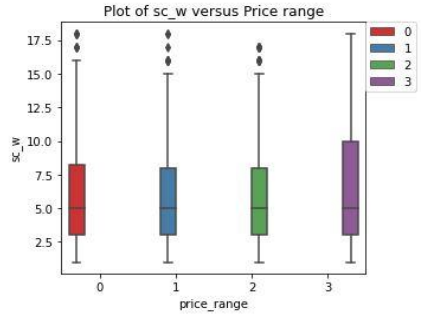
Effect of Target variable on features:

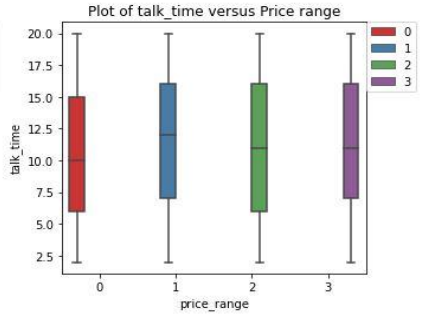








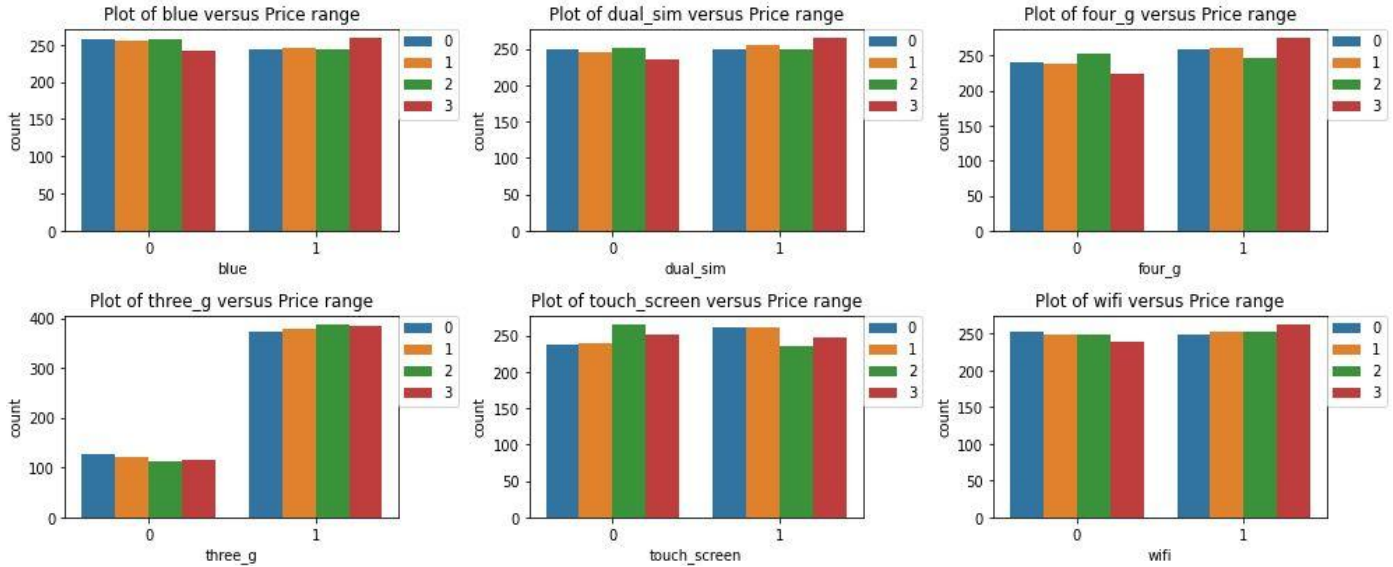




EDA

Bivariate Analysis contd.

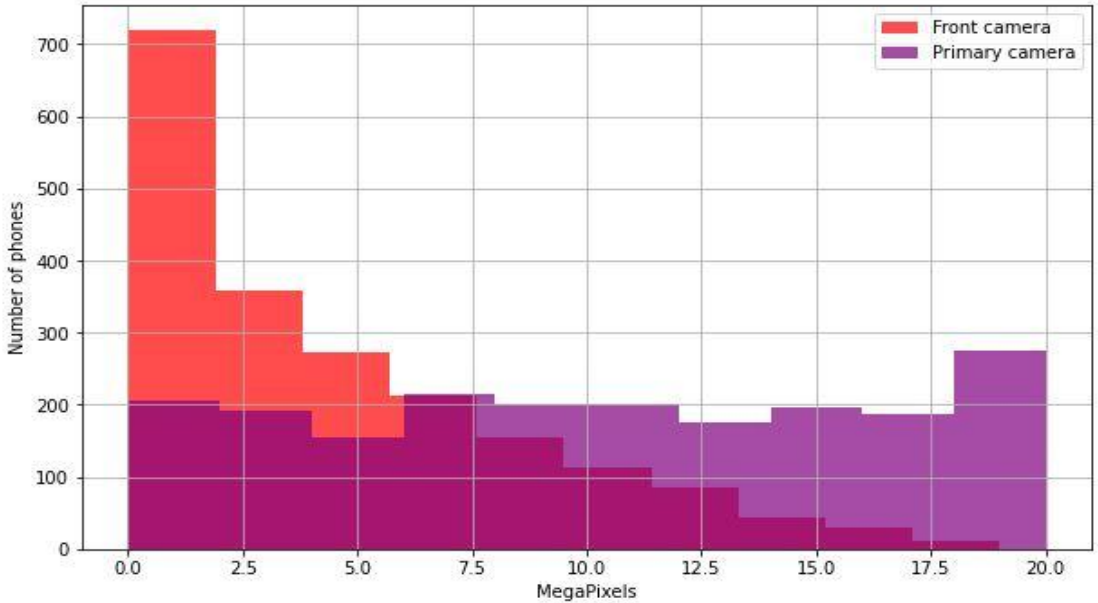
Effect of Target variable on features:



EDA

Multivariate Analysis

Effect of Primary and front camera on phones:



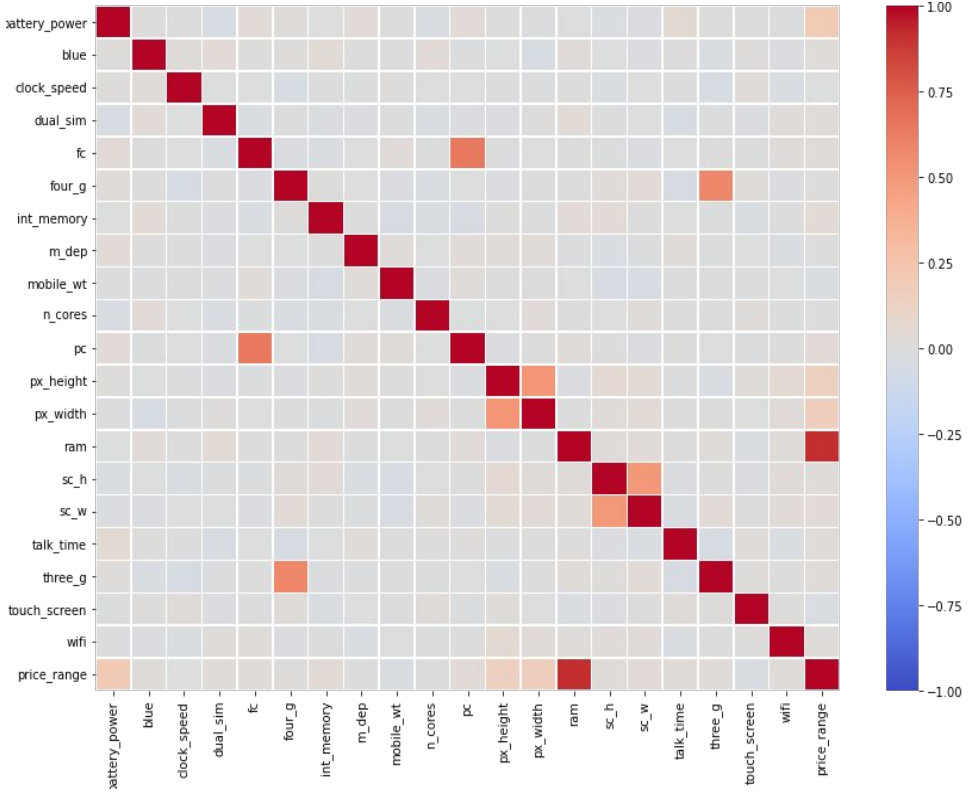
EDA

Conclusion:

- **Relatively expensive phones have higher capacity batteries.**
- **Most Expensive (category 3) phones have better cameras.**
- **Relatively expensive phones have much bigger RAMs.**
- **Expensive phones have better screen quality.**
- **Most Expensive phones have wider screens.**
- **Most Expensive phones are lighter than competition.**

Feature Engineering

Checking for multicollinearity:



Feature Engineering contd.

Checking for multicollinearity using VIF score:

	feature	VIF
0	battery_power	8.073908
1	blue	1.982736
2	clock_speed	4.261359
3	dual_sim	2.014356
4	fc	3.412883
5	four_g	3.193731
6	int_memory	3.960540
7	m_dep	3.910499
8	mobile_wt	12.969326
9	n_cores	4.647420
10	pc	6.222849
11	px_height	4.280154
12	px_width	11.792757
13	ram	4.688630
14	sc_h	11.675539
15	sc_w	4.650102
16	talk_time	4.857407
17	three_g	6.195323
18	touch_screen	1.988861
19	wifi	2.020265

Feature Engineering contd.

Using SelectKbest method with ‘chi2’ as scoring function for feature selection:

	features	scores
0	battery_power	14129.866576
1	blue	0.723232
2	clock_speed	0.648366
3	dual_sim	0.631011
4	fc	10.135166
5	four_g	1.521572
6	int_memory	89.839124
7	m_dep	0.745820
8	mobile_wt	95.972863
9	n_cores	9.097556
10	pc	9.186054
11	px_height	17561.692129
12	px_width	9810.586750
13	ram	931267.519053
14	sc_h	9.614878
15	sc_w	10.764356
16	talk_time	13.236400
17	three_g	0.327643
18	touch_screen	1.928429
19	wifi	0.422091

features	scores
ram	931267.519053
px_height	17561.692129
battery_power	14129.866576
px_width	9810.586750
mobile_wt	95.972863
int_memory	89.839124
talk_time	13.236400
sc_w	10.764356
fc	10.135166
sc_h	9.614878
pc	9.186054
n_cores	9.097556

Model Implementation

Following Models were implemented:

- **Logistic Regression**
- **K Nearest Neighbors**
- **Support Vector Machine**
- **Naive Bayes Classifier**
- **Random Forest**
- **XGBoost**

Model Implementation contd.

	Classification Model	Accuracy	Recall	Precision	F1-score
0	Logistic (Baseline)	0.938	0.938	0.937537	0.937490
1	Logistic (Tuned)	0.968	0.968	0.968194	0.968017
2	KNN (Baseline)	0.614	0.614	0.641735	0.623575
3	KNN (Tuned)	0.790	0.790	0.801598	0.792836
4	SVM (Baseline)	0.948	0.948	0.948425	0.948119
5	SVM (Tuned)	0.968	0.968	0.968129	0.967984
6	Naive Bayes	0.834	0.834	0.834042	0.834015
7	Random Forest (Baseline)	0.888	0.888	0.887766	0.887826
8	Random Forest (Tuned)	0.896	0.896	0.894624	0.895094
9	XGBoost (Baseline)	0.906	0.906	0.905800	0.905878
10	XGBoost (Tuned)	0.928	0.928	0.927421	0.927524

Model Selection

Observations:

- KNN is the worst performing algorithm even after tuning.
- Naive Bayes is the second worst algorithm here.
- Logistic Regression, SVM, XGBoost gained the most from tuning.
- Logistic Regression, SVM are the best performing models getting accuracy of 96.8% (after tuning), followed by XGBoost having accuracy 92.8% (after tuning).

Model Selection contd.

Logistic Regression

Accuracy score is 0.968

Classification report

	precision	recall	f1-score	support
0	0.976	0.984	0.980	124
1	0.938	0.964	0.951	110
2	0.967	0.944	0.955	125
3	0.986	0.979	0.982	141
accuracy			0.968	500
macro avg	0.967	0.968	0.967	500
weighted avg	0.968	0.968	0.968	500

SVM

Evaluating the performance on Test data
Accuracy score is 0.968

Classification report

	precision	recall	f1-score	support
0	0.984	0.992	0.988	124
1	0.938	0.964	0.951	110
2	0.959	0.936	0.947	125
3	0.986	0.979	0.982	141
accuracy			0.968	500
macro avg	0.967	0.968	0.967	500
weighted avg	0.968	0.968	0.968	500

XGBoost

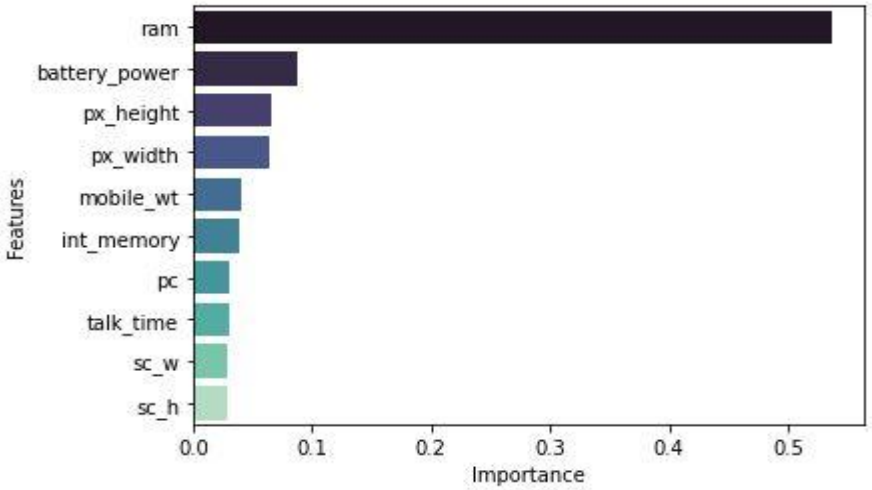
Evaluating the performance on Test data
Accuracy score is 0.928

Classification report

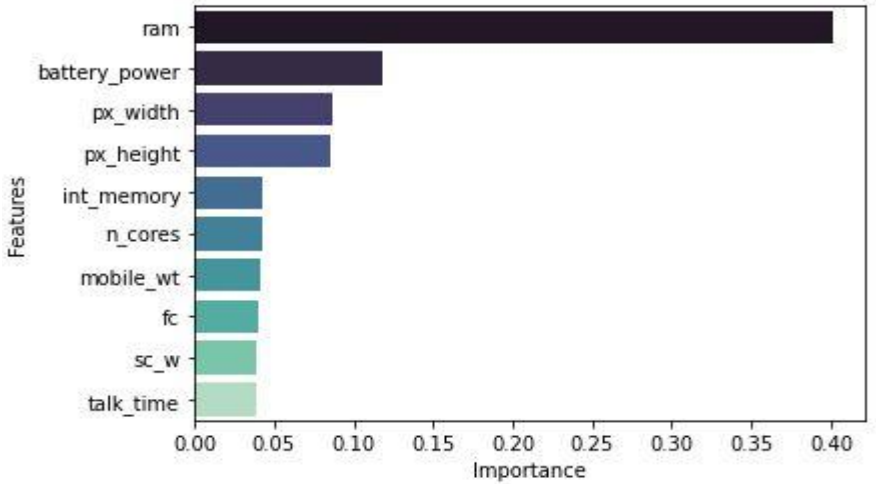
	precision	recall	f1-score	support
0	0.961	0.992	0.976	124
1	0.925	0.900	0.912	110
2	0.893	0.864	0.878	125
3	0.931	0.950	0.940	141
accuracy			0.928	500
macro avg	0.927	0.927	0.927	500
weighted avg	0.927	0.928	0.928	500

Feature Importance

Random Forest



XGBoost



Conclusion

- After evaluating the performance metrics of all the models, we have come to the conclusion that SVM and Logistic Regression are the best models for this case.
- SVM and Logistic Regression scored accuracy of 96.8% on the test set.
- KNN was the worst performing model with accuracy of 79%
- The important features in determining price range of a mobile phone were: RAM, battery power, pixel height and pixel width.
- But the most important feature which single handedly affects the price range is RAM as can be seen from the selectkbest score and feature important graphs.

Thank you