Capstone Project-3

Bike Sharing demand prediction

Team Members:

Sharad Tawade
Sagar Malik
Vinay Kumar

Contents:

Need for Mobile Range Prediction?

- Mobile is a very competitive market segment that cannot be easily captured. Margin in the technology used is less, which is compensated by the volumes sold.
- Avg. mobile sold in a year are around 1500 millions, which will keep growing in the coming years as economic sector improves more and more people will be able to afford mobile phones also the increasing population.
- Looking at the above two points we see the there is a need for segregating mobile phones into various price categories.
- Now for a company that want to develop a mobile with some specifications and wanna know the budget for each unit that will be made.
- This is the place where this model will come into clutch.

Introduction

- Mobile phones are now like an essential commodity for us to connect to the world or our loved ones, resulting in the huge amount of mobile phone manufactured, hence; huge amount of data being generated.

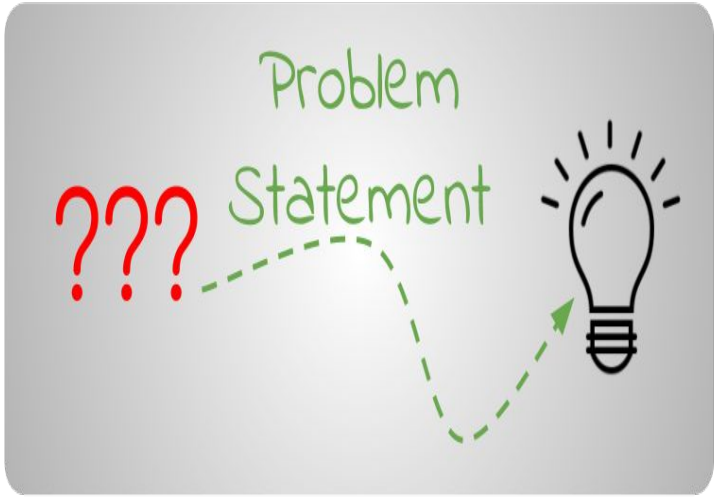
- Mobile phone prediction helps in deciding the range of a mobile phone depending upon its specifications, as the most expensive mobile phone will be loaded with a lot more and better features than the cheap ones.

- This insight can help deciding the specification for a mobile phone at industry level.

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.



Description of the dataset!

- **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 value of 0(low cost), 1(medium cost), 2(high cost) and 3(very high cost).

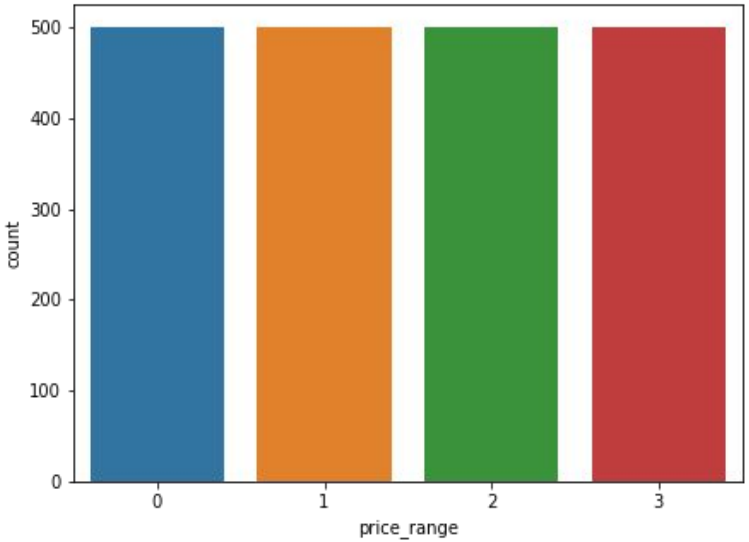
Data wrangling

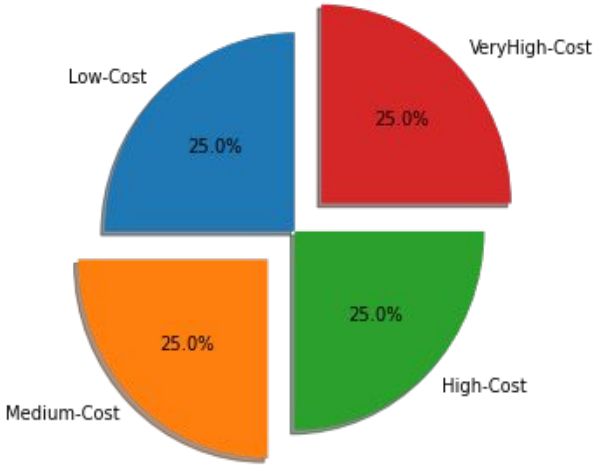
- **Outlier Detection** : No outlier found in categorical as well as numerical features.
- No Null values found
- No data Transformation needed.

EXPLORATORY DATA ANALYSIS

Analysing price range

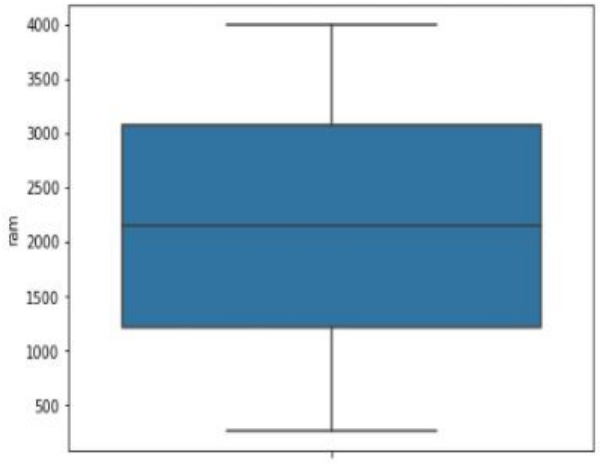
- We have a balanced dependent variable, where we have 500 mobiles lying in each category (0,1,2,3).

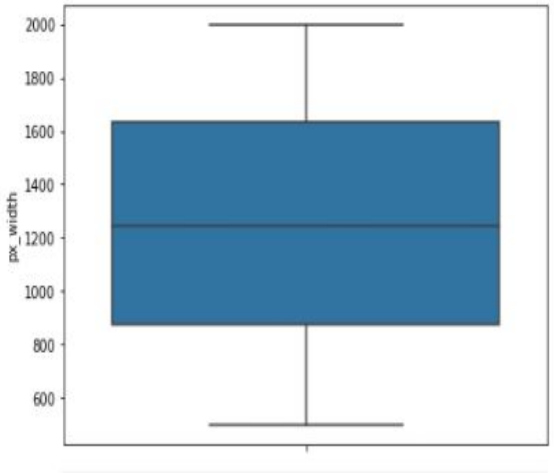


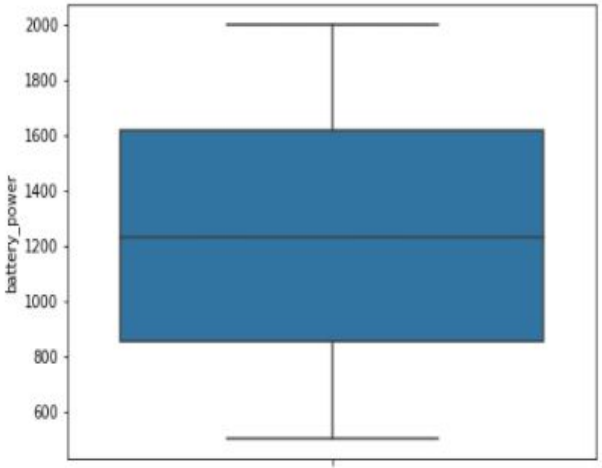


Check for outliers

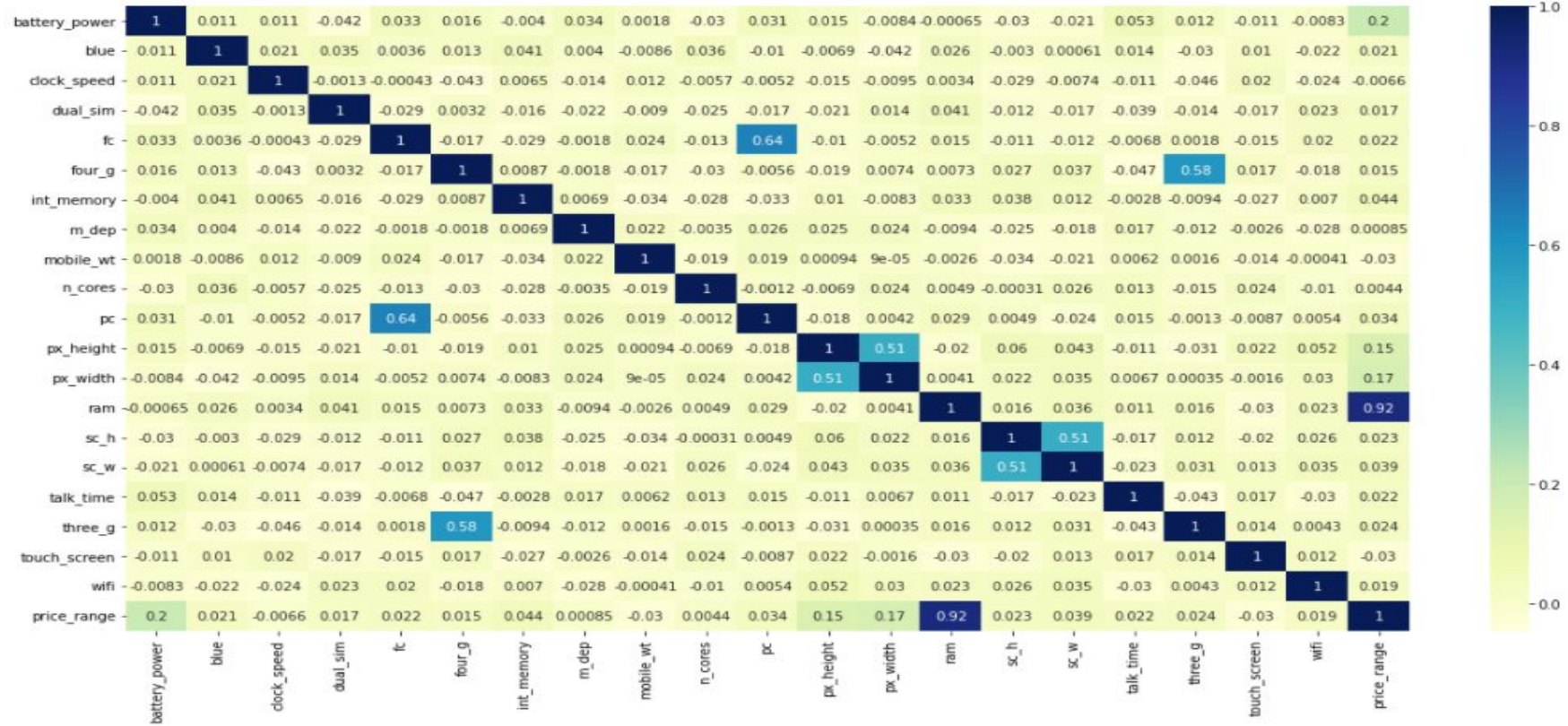
- No problem related to outliers were faced, it was a very well balanced dataset. Below are the most important features and as we can clearly see, there are no outliers.



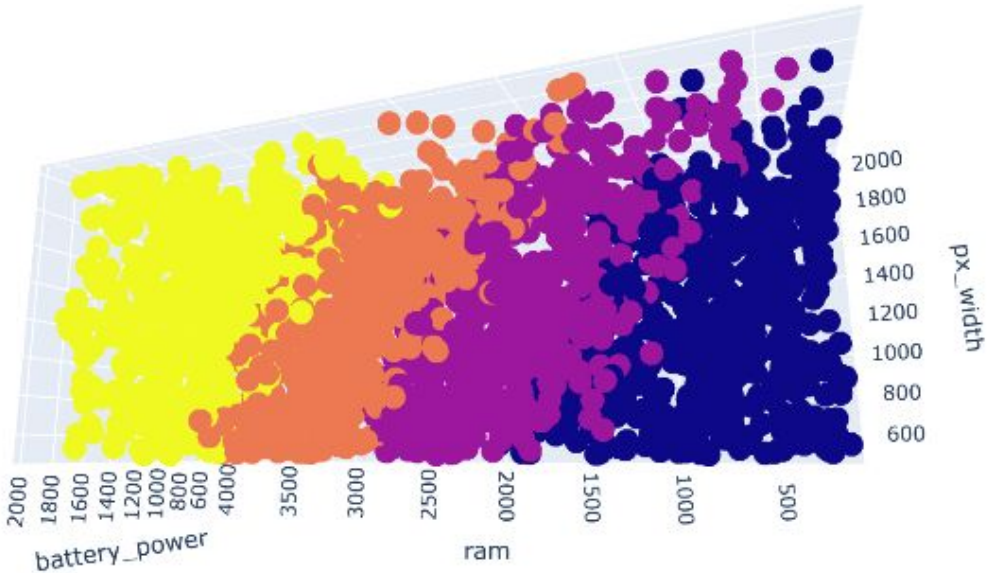




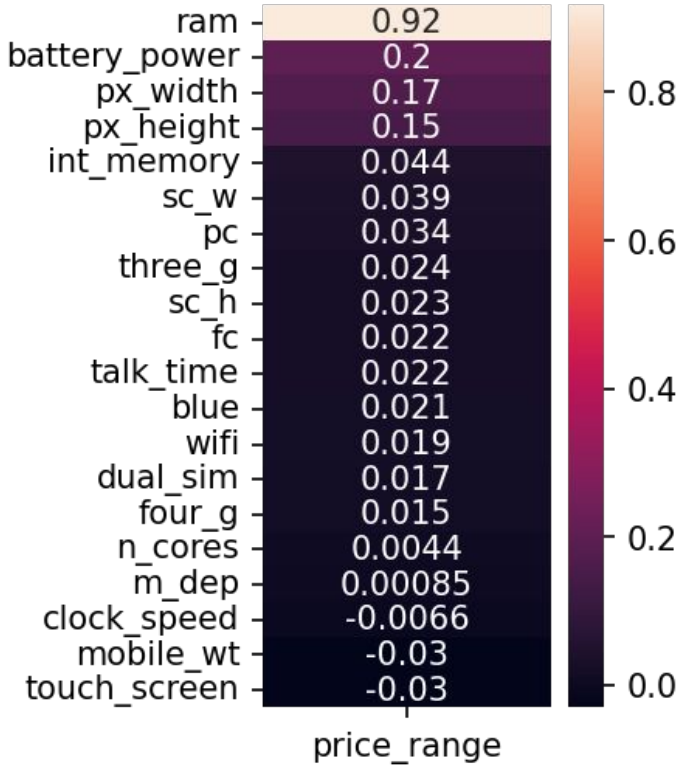
Multicollinearity



Multicollinearity (conti.)



- 3D correlation plot between ram, px_width & battery_power.



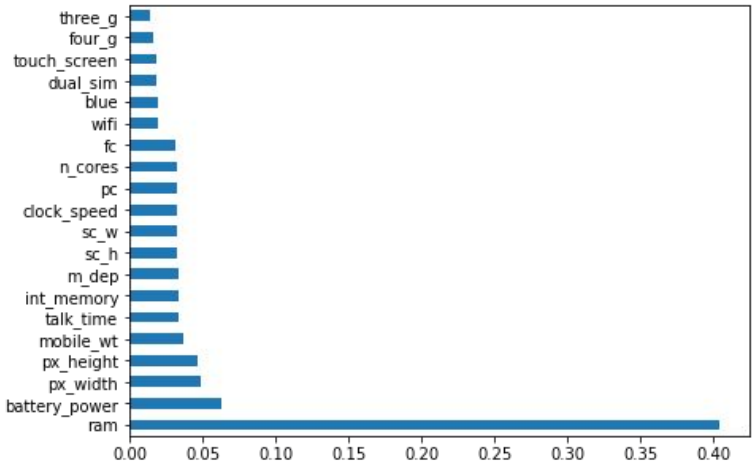
- Ram is the most correlated to price .i.e., it has the most impact on the price of a mobile

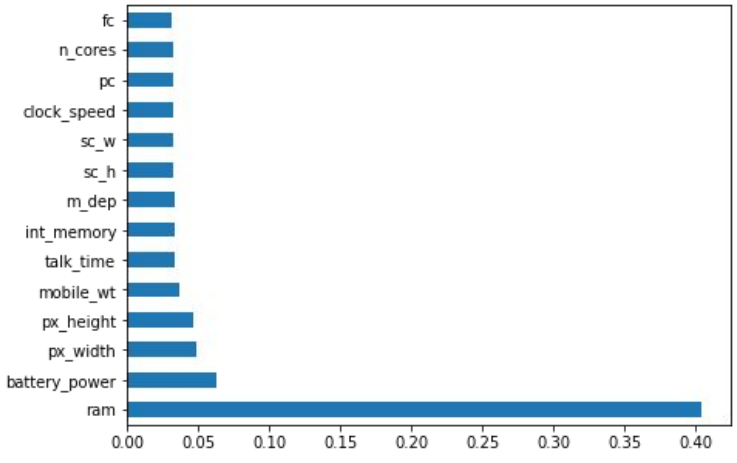
Feature Selection:

In Feature selection we remove non-informative features from the dataset. At beginning we have 2000 rows and 21 columns. After feature engineering we get 2000 rows and 14 columns.

Method for Feature Selection

- we will use ExtraTreesClassifier from the sklearn library for Feature Importance Selection.
- This class implements a meta estimator that fits a number of randomized decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.
- RAM being the most influential and all the categorical columns being the least contributor to the price of a mobile.
- Thus; selecting top 14 features for further processes.





Model’s Performed

- Logistic Regression
- Decision Tree
- Random Forest
- XGBoost
- K-Nearest Neighbour
- SVC
- Neural Network

Applying Models - Validation & Selection

	Model_Name	Training_accuracy	Test_accuracy	Train_roc_auc_score	Test_roc_auc_score
0	Logistic Regression	0.978	0.955	0.999	0.998
1	Decision Tree	0.941	0.855	0.996	0.944
2	KNN	0.731	0.635	0.907	0.868
3	Support Vector Machine	0.980	0.917	0.999	0.990
4	Neural-Network	0.983	0.975	0.999	0.998

We remove 'Random Forest' & 'XGBoost' from our model comparison because they were overfitting the data, also we had other models which working perfectly so that we don't need them in the comparison.

Confusion Matrix of the Models

Logistic Regression

```
The confusion matrix on the train data is :
[[403   7   0   0]
 [  2 395  10   0]
 [  0   6 384   3]
 [  0   0   7 383]]
```

```
The confusion matrix on the train data is :
[[ 93   2   0   0]
 [  2  88   5   0]
 [  0   2  90   3]
 [  0   0   4 111]]
```

KNN

```
The confusion matrix on the train data is :
[[346  73   4   0]
 [ 58 266  84   6]
 [  1  65 271  92]
 [  0   4  42 288]]
```

```
The confusion matrix on the train data is :
[[76 25  0  0]
 [18 46 32  3]
 [ 1 20 59 38]
 [ 0  1  8 73]]
```

Neural Networks

```
The confusion matrix on the train data is :
[[407   1   0   0]
 [  0 389   2   0]
 [  0   3 387   3]
 [  0   0   4 404]]
```

```
The confusion matrix on the train data is :
[[ 92   2   0   0]
 [  1 100   3   0]
 [  0   5 103   1]
 [  0   0   1  92]]
```

(Best Model)

Decision Tree

```
The confusion matrix on the train data is :
[[386  18   0   0]
 [ 19 372  18   0]
 [  0  18 361  15]
 [  0   0  22 371]]
```

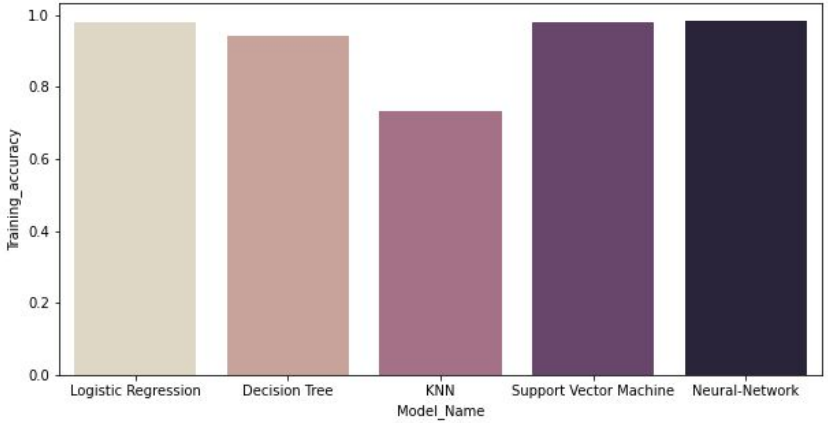
```
The confusion matrix on the train data is :
[[ 86   6   0   0]
 [  9  71   8   0]
 [  0  15  81  11]
 [  0   0  10 103]]
```

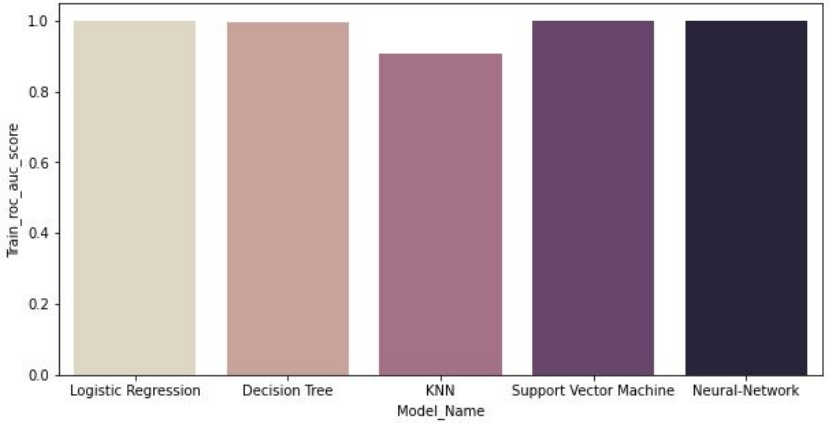
SVC

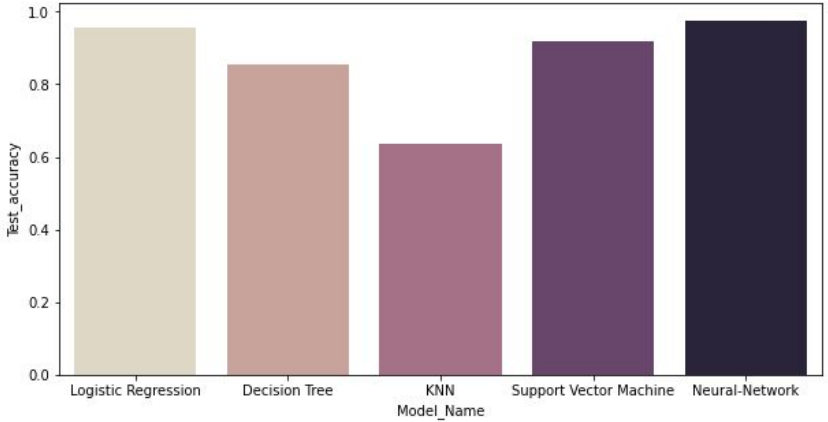
```
The confusion matrix on the train data is :
[[400   3   0   0]
 [  5 397   7   0]
 [  0   8 393   5]
 [  0   0   1 381]]
```

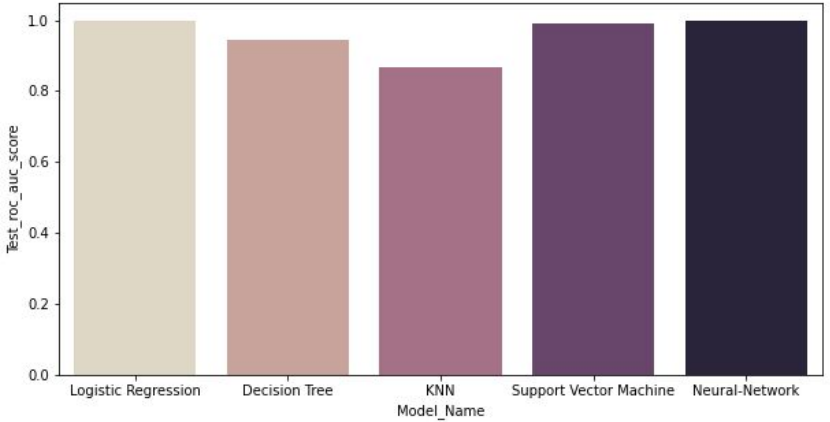
```
The confusion matrix on the train data is :
[[ 92   4   0   0]
 [  3  82   7   0]
 [  0   6  89  10]
 [  0   0   3 104]]
```

Model - Validation & Selection









Model - Validation & Selection

Observation 1: As observed from the table above Logistic Regression and decision tree model performs well along with support vector machine and neural network model.

Observation 2: Random forest & X-Gradient Boost model may overfit the model.

Observation 3: From the above observation we have come to a conclusion that we would choose our classification model from Neural Network.

Difficulties Faced

- Being a balanced dataset, the difficulties were on a lower level.
- Some models overfitted the data.
- Feature selection.
- Model validation as most of the algorithm were performing almost similar.

Conclusion

- As observed ***Logistic Regression*** and ***decision tree*** model performs well along with ***support vector machine*** and ***neural network model***.
- Thus; for production we will go with our ***Neural Network Model.***
- ***Ram***, ***Battery power***, ***Mobile weight***, ***Screen size*** and ***Pixels*** are key features in predicting the mobile price range.

THANK YOU