Capstone Project
Mobile price range prediction

Nitesh Bhowmick

Content

Problem statement

- Prediction based on Mobile price.
- Prediction based on Bluetooth connectivity .
- Prediction based on Battery with price.
- Prediction based on the RAM.
- Prediction based on the weight.
- Prediction based on the Battery power, pixels played etc.

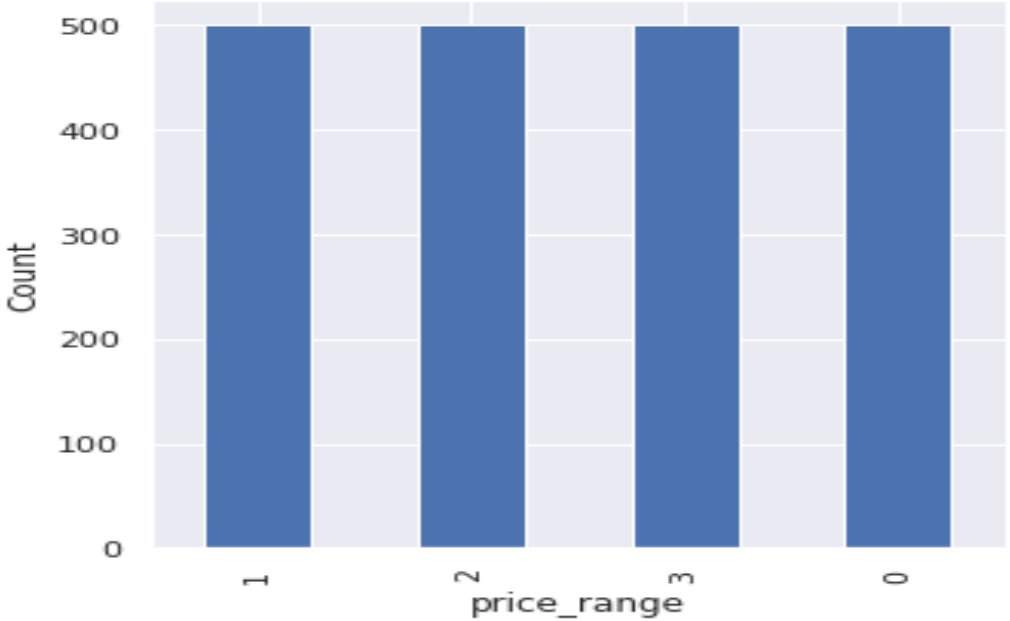
Data summary

- In the Mobile price Range Prediction project there is a Dataset which contains mobile information
- The dataset has total of 21 columns and 2000 rows
- The Dataset contains information on mobile price, ram , Bluetooth connectivity, Battery power,3G and 4G.

Exploratory Data Prediction

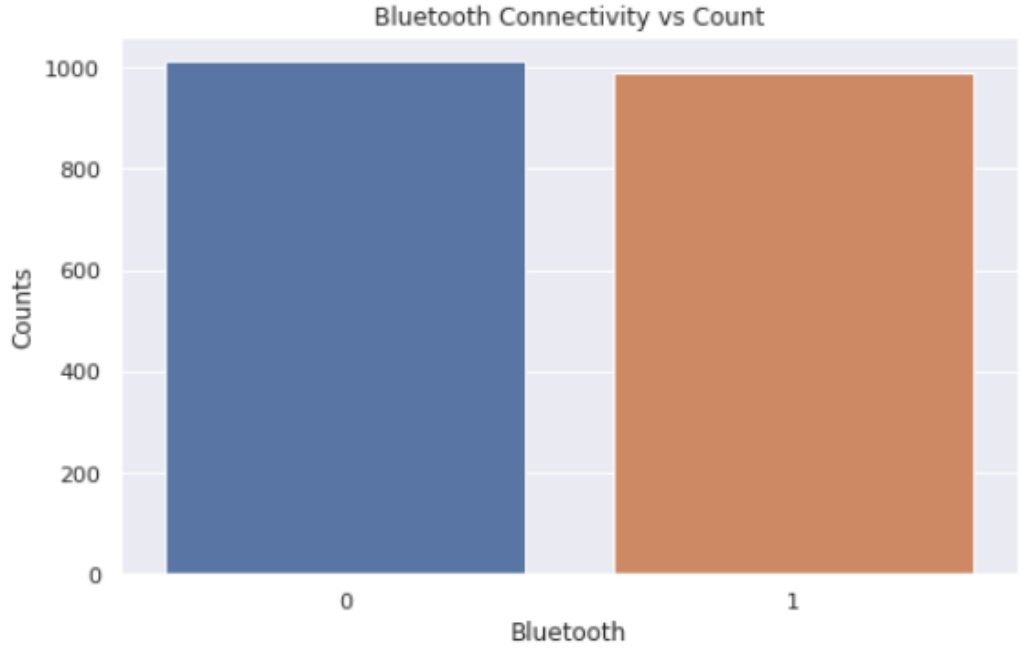
- Exploratory Data Prediction is also known as EDA, is the process of interpreting datasets by summarizing their key properties and frequently visualizing them
- EDA refers to the critical process of performing initial investigation on dataset so as discover the patterns , to spot anomalies, to hypothesis, and to check assumptions with the help of summary statistics and graphical representation.
- In EDA , plotting option include box plots, line plots , scatter plots and many more

EDA For Mobile price range Prediction



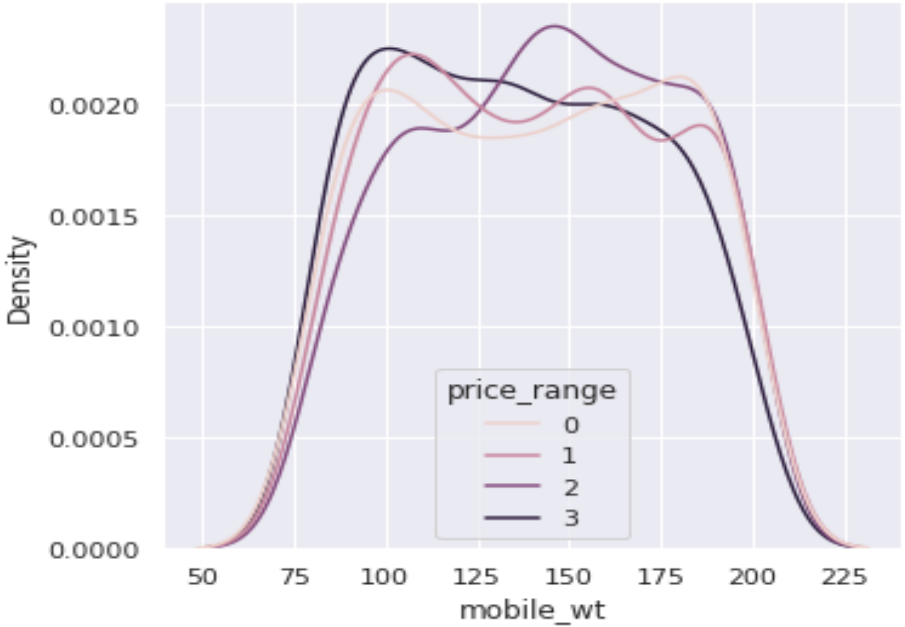
- There are mobile phone in 4 price range . The number of elements is almost similar.

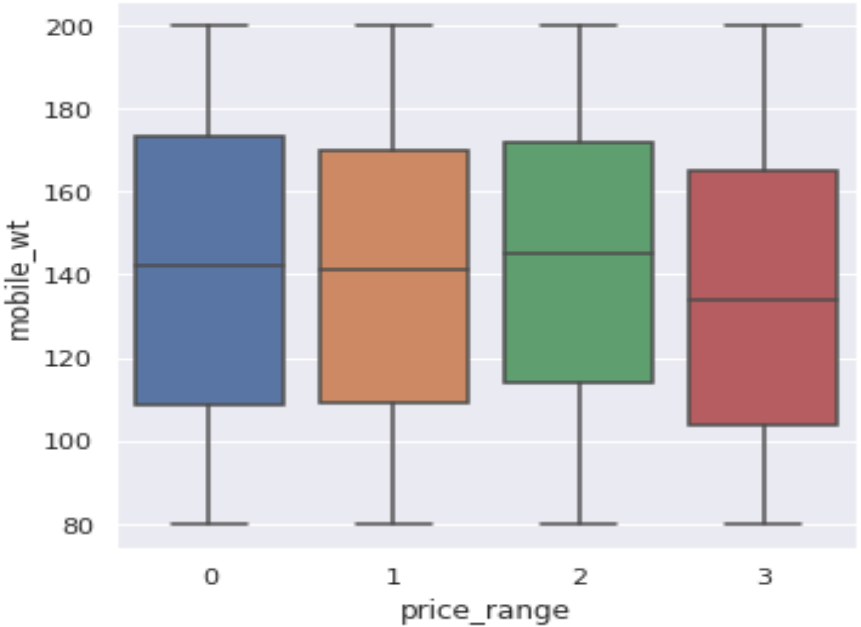
Bluetooth connectivity



As, we see approximately half of devices have bluetooth connectivity & another half of devices don’t have bluetooth connectivity.

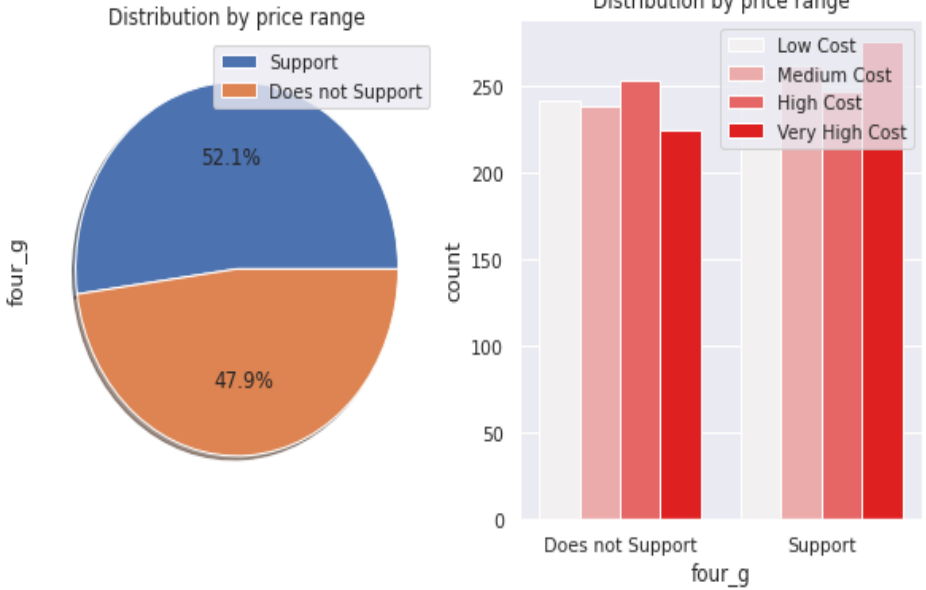
Mobile weight

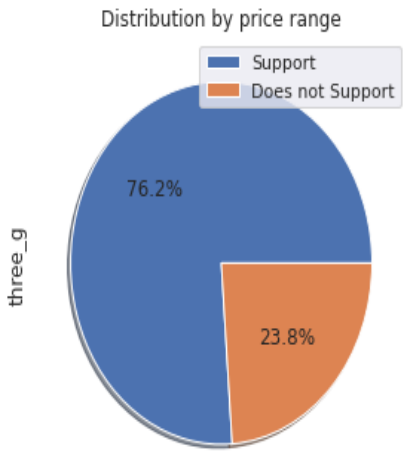


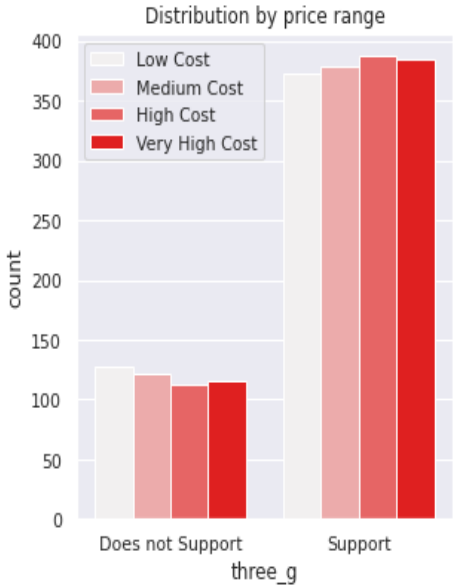


As we can see , costly phones are lighter.

Distribution by price range for Connectivity feature

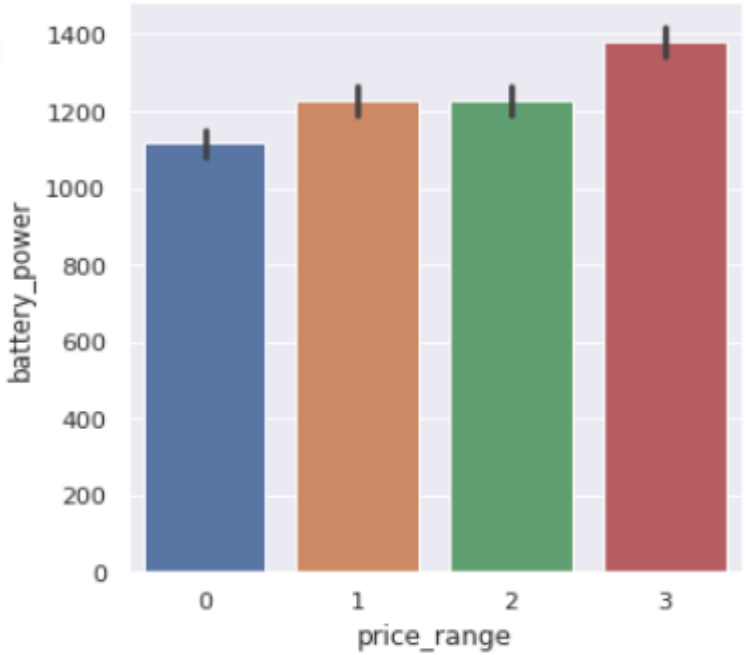


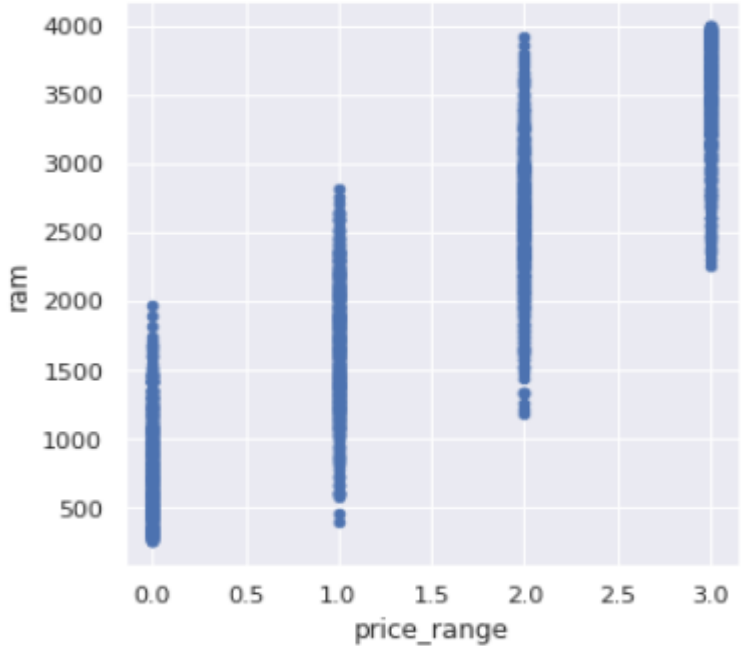




- Connectivity feature 3G play an important feature than 4G in prediction.

Battery power & Ram with price Range





This plots shows how the battery mAh is spread. there is a gradual increase as the price range increases &

Ram has continuous increase with price range while moving from Low cost to Very high cost.

Machine learning

- Machine learning is a type of artificial intelligence that allows software application to become more accurate at predicting outcomes without being explicitly programmed to do so machine learning algorithms use historical data as input to predict new output value
- Machine learning are a common use case for machine learning Other popular uses include fraud detection, business process automation (BPA) and predictive maintenance.

Why machine learning is important

- Machine learning is important because it gives enterprises a view of trends in customers behavior and business operational patterns, as well as supports the development of new products . Many of today , leading companies such as Facebook, Google and Uber, make machine learning a central part of their operations. Machine learning has become a significant competitive differentiator for many companies.

Different types of machine learning

- **Supervised learning:** In this types of machine learning, data scientists supply algorithms with labeled training data and defined the variables they want the algorithm to assess for correlation.
- **Unsupervised Learning:** This types of machine learning involves algorithms that train on unlabeled data.
- **Reinforcement learning:** It is multi step process for which there are clearly defined rules Data scientist program an algorithm complete a task.

Logistic Regression

Logistic regression : Logistic regression is a classification algorithm that predicts the probability of an outcome that can only have two values (i.e. a dichotomy). A logistic regression produces a logistic curve, which is limited to values between 0 and 1. Logistic regression models the probability that each input belongs to a particular category.

Advantages and disadvantages of logistic regression:-

Logistic regression is easier to implement, interpret, and very efficient to train. If the number of observations is lesser than the number of features, Logistic Regression should not be used, otherwise, it may lead to overfitting.

Import Logistic Regression:-

```
from sklearn.linear_model import LogisticRegression
```

Random Forest

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

Advantages of Random Forest:-

- Random Forest is capable of performing both Classification and Regression tasks.
- It is capable of handling large datasets with high dimensionality.
- It enhances the accuracy of the model and prevents the over fitting issue.

Disadvantages of Random Forest:-

- Although random forest can be used for both classification and regression tasks, it is not more suitable for Regression tasks.

Import Random Forest Classifier:-

- from sklearn.ensemble import RandomForestClassifier

Decision tree

Decision Tree is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label.

Advantages and disadvantages of the decision tree:-

- They are very fast and efficient compared to KNN and other classification algorithms. Easy to understand, interpret, visualize. The data type of decision tree can handle any type of data whether it is numerical, categorical or Boolean. Normalization is not required in the Decision Tree.
- One of the limitations of decision trees is that they are largely unstable compared to other decision predictors. A small change in the data can result in a major change in the structure of the decision tree, which can convey a different result from what users will get in a normal event.

Import Decision Tree Classifier:-

- from sklearn.tree import DecisionTreeClassifier

Support vector Machine

AI

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes .This best decision boundary is called a hyperplane.
SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine.

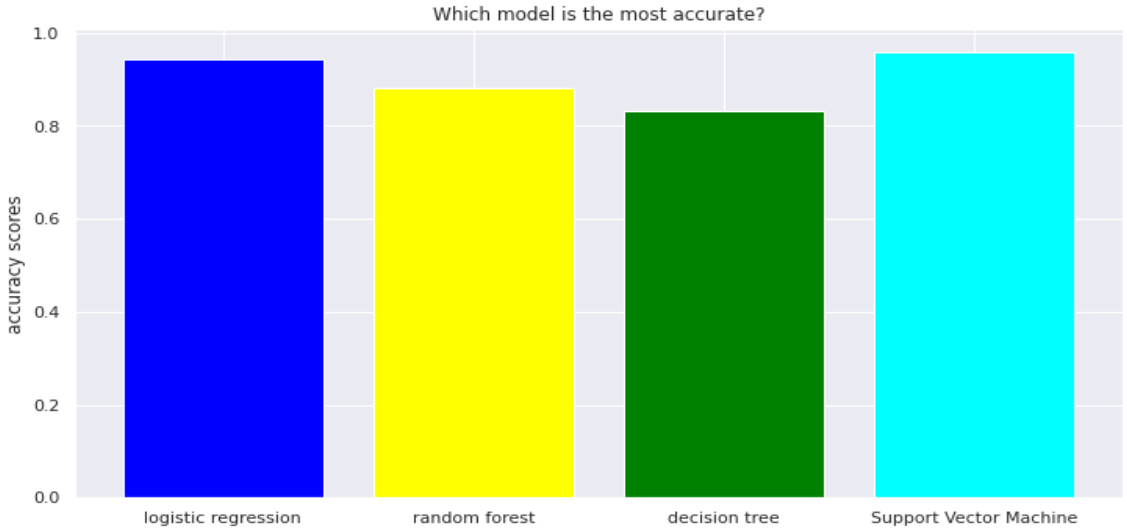
Advantages and disadvantages of the Support Vector Machine:-

- Support vector machine works comparably well when there is an understandable margin of dissociation between classes. It is more productive in high dimensional spaces. It is effective in instances where the number of dimensions is larger than the number of specimens.
- It does not execute very well when the data set has more sound i.e. target classes are overlapping. In cases where the number of properties for each data point outstrips the number of training data specimens, the support vector machine will underperform.

Import Support Vector Classifier:-

```
from sklearn import svm
from sklearn.svm import SVC
```

Accuracy Scores



- After training our dataset with four different model, we conclude that logistic regression & svm are the best model for our dataset.

Challenges

- Understand the column of the dataset.
- Analyze and visualization of the price range according features.
- Find the right chart to show the chart.
- Difficulties to find out the correct model.

Conclusion

- In EDA we can see that there are mobile phone in 4 price range. The no of elements is almost similar.
- Half of devices have Bluetooth connectivity & another half of devices don’t have Bluetooth connectivity.
- There is a gradual increase in battery as the price range increases.
- Costly phone are lighter.
- Ram , battery power, pixels and connectivity feature 3G & 4G more significant role in deciding the price range phone.
- Deploy 4 Machine learning model in our dataset.
- From all the above experiment we can conclude that logistic regression and SVM with hyper parameters we got the best accuracy score.

Q/A