Mobile Price Prediction

Almabetter, Bangalore

Neha Pasi

Abstract:

Mobile phones come in all sorts of prices, features, specifications and all. Price estimation and prediction is an important part of consumer strategy. Deciding on the correct price of a product is very important for the market success of a product. A new product that has to be launched, must have the correct price so that consumers find it appropriate to buy the product.

Problem Statement:

The data contains information regarding mobile phone features, specifications etc and their price range. The various features and information can be used to predict the price range of a mobile phone.

Introduction to EDA:

Exploratory Data Analysis is investigating data and drawing out insights from it to study its main characteristics. EDA can be done using statistical and visualization techniques.

Exploring and analyzing the data is important to see how features are contributing to the target variable, identifying anomalies and outliers to treat them lest they affect our model, to study the nature of the features, and be able to perform data cleaning so that our model building process is as efficient as possible.

If we don’t perform exploratory data analysis, we won’t be able to find inconsistent or incomplete data that may pose trends incorrectly to our model.

This step also serves as the basis for answering our business questions.

Introduction to Supervised Classification Machine learning:

In supervised learning, you train your model on a labelled dataset that means we have both raw input data as well as its results. We split our data into a training dataset and test dataset where the training dataset is used to train our network whereas the test dataset acts as new data for predicting results or to see the accuracy of our model.

Hence, in supervised learning, our model learns from seen results the same as a teacher teaches his students because the teacher already knows the results. Accuracy is what we achieve in supervised learning as model perfection is usually high.

The model performs fast because the training time taken is less as we already have desired results in our dataset. This model predicts accurate results on unseen data or new data without even knowing a prior target. In some of the supervised learning models, we revert back the output result to learn more in order to achieve the highest possible accuracy.

Data Summary:

Based upon the initial assessment we found that the data was pretty much clean except for Outliers in some columns. We draw out the following key insights about the data:-

1. The dataset has a shape of (2000, 21) which means that it contains approximately 2000 rows and 21 columns.

The data features are as follows:

- 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).

Methodology

We will proceed with reading the data, and then perform data analysis. The practice of examining data using analytical or statistical methods in order to identify meaningful information is known as data analysis. After data analysis, we will find out the data distribution and data types. We will train 4 classification algorithms to predict the output. We will also compare the output.

Steps Involved:

- Exploratory Data Analysis

  After loading the dataset we performed this method by comparing our target variable that is Trip duration with other independent variables. This process helped us figuring out various aspects and relationships among the target and the independent variables. It gave us a better idea of which feature behaves in which manner compared to the target variable.

- Null Values Treatment

  We don’t have any null values in our dataset.

- Independent and Dependent Variable Selection

  In this step we select appropriate dependent and and independent variable for making prediction.

- Split the Dataset into Train and Test Sets

  We need to split a data set into train and test sets to evaluate how well our machine learning model performs. The train set is used to fit the model. The second set is called the test data set, this set is solely used for predictions.

- Fitting different Models

  For modeling we tried various Classification algorithms like:

  1. Random Forest Algorithm
  2. Naive Bayes
  3. KNN Classifier
  4. SVM(support vector machine) Algorithm

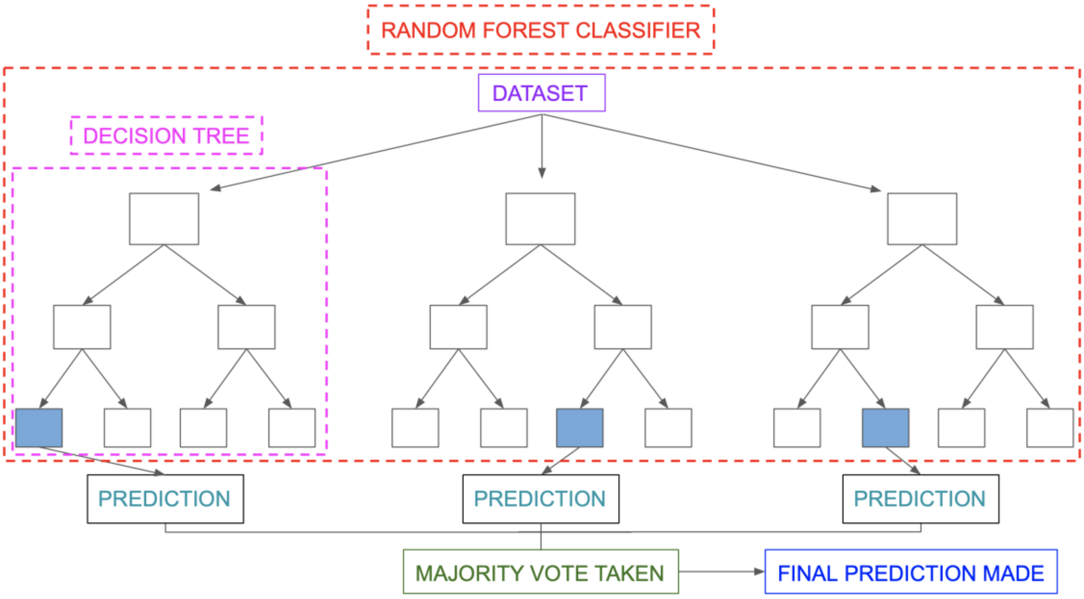
Random Forest Classifier

A random forest is a supervised machine learning method built from decision tree techniques. This algorithm is used to anticipate behaviour and results in a variety of sectors, including banking and e-commerce.

A random forest is a machine learning approach for solving regression and classification issues. It makes use of ensemble learning, which is a technique that combines multiple classifiers to solve complicated problems.

A random forest method is made up of a large number of decision trees. The random forest algorithm's 'forest' is trained via bagging or bootstrap aggregation. Bagging is a meta-algorithm ensemble that increases the accuracy of machine learning algorithms.

The outcome is determined by the (random forest) algorithm based on the predictions of the decision trees. It forecasts by averaging or averaging the output of several trees. The precision of the outcome improves as the number of trees grows.



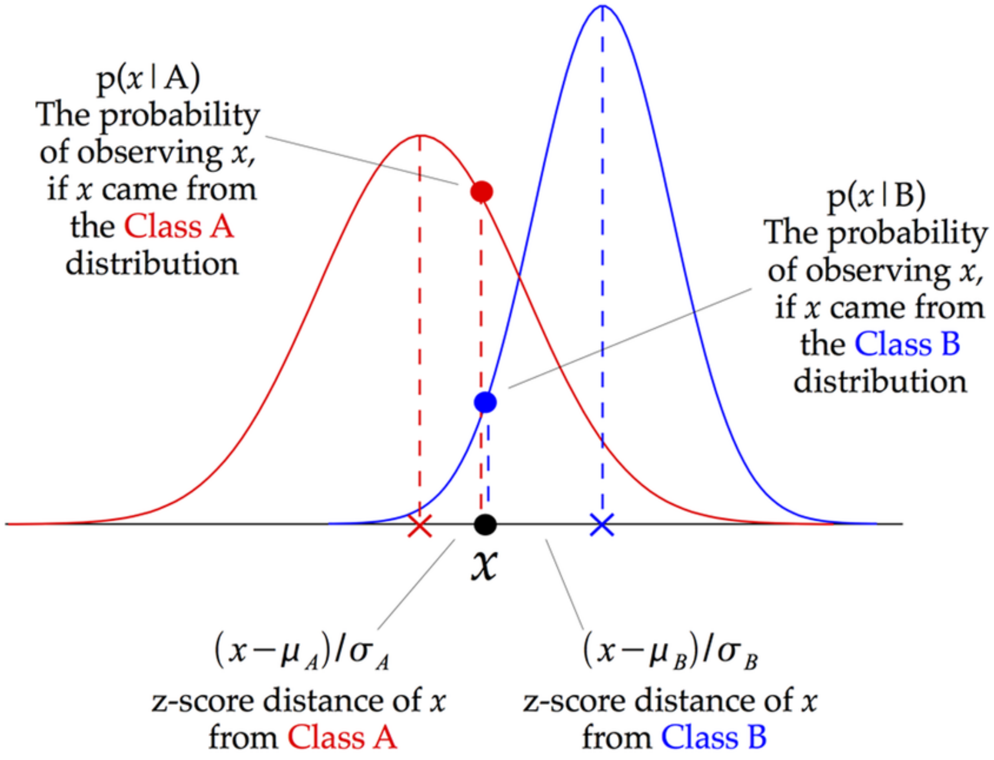
Image

Source: https://miro.medium.com/max/5752/1*5dq_1hnqkboZTcKFfwbO9A.png

A random forest system is built on a variety of decision trees. Every decision tree is made up of nodes that represent decisions, leaf nodes, and a root node. The leaf node of each tree represents the decision tree's final result. The final product is chosen using a majority-voting procedure. In this situation, the output picked by the majority of the decision trees becomes the random forest system's ultimate output. Let us now implement the random forest algorithm.

Naive Bayes

Conditional probability is the foundation of Bayes' theorem. The conditional probability aids us in assessing the likelihood of something occurring if something else has previously occurred.



$p(x \mid A)$
The probability of observing x, if x came from the Class A distribution
$p(x \mid B)$
The probability of observing x, if x came from the Class B distribution
x
$(x-\mu_A)/\sigma_A$
z-score distance of x from Class A
$(x-\mu_B)/\sigma_B$
z-score distance of x from Class B

Image: Illustration of how a Gaussian Naive Bayes (GNB) classifier works

Source: https://www.researchgate.net/figure/Illustration-of-how-a-Gaussian-Naive-Bayes-GNB-classifier-works-For-each-data-point_fig8_255695722

Gaussian Naive Bayes is a Naive Bayes variation that allows continuous data and follows the Gaussian normal distribution. The Bayes theorem is the foundation of a family of supervised machine learning classification algorithms known as naive Bayes. It is a basic categorization approach with a lot of power. When the dimensionality of the inputs is high, they are useful. The Naive Bayes Classifier may also be used to solve complex classification issues.

KNN Classifier

The K Nearest Neighbor method is a type of supervised learning technique that is used for classification and regression. It’s a flexible approach that may also be used to fill in missing values and resample datasets. K Nearest Neighbor examines K Nearest Neighbors (Data points) to forecast the class or continuous value for a new Datapoint, as the name indicates.

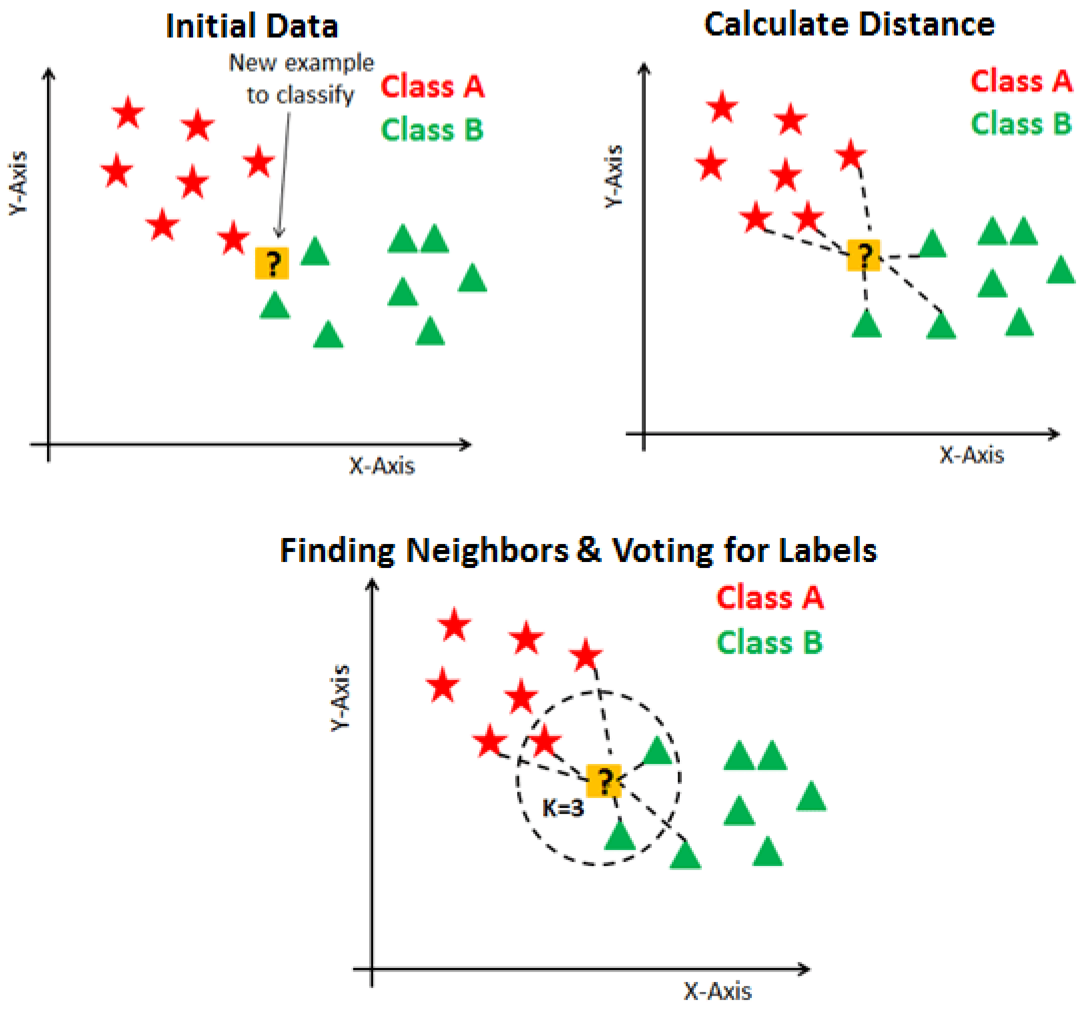


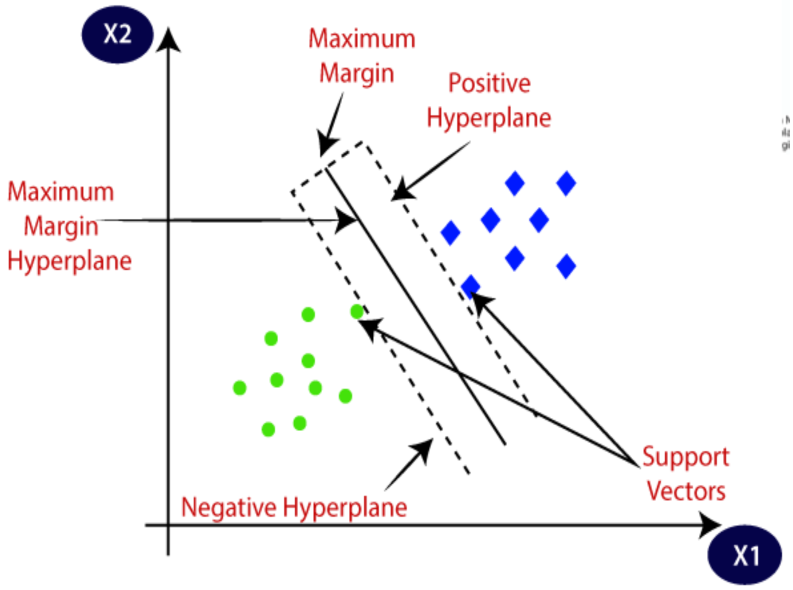
Image: https://blakelobato1.medium.com/k-nearest-neighbor-classifier-implement-homemade-class-compare-with-sklearn-import-6896f49b89e

The K-NN method saves all available data and classifies a new data point based on its similarity to the existing data. This implies that fresh data may be quickly sorted into a well-defined category using the K-NN method. The K-NN algorithm is a non-parametric algorithm, which means it makes no assumptions about the underlying data. It’s also known as a lazy learner algorithm since it doesn’t learn from the training set right away; instead, it saves the dataset and performs an action on it when it comes time to classify it.

SVM Classifier

Support Vector Machine, or SVM, is a prominent Supervised Learning technique that is used for both classification and regression issues. However, it is mostly utilised in Machine Learning for Classification purposes.

The SVM algorithm’s purpose is to find the optimum line or decision boundary for categorising n-dimensional space so that we may simply place fresh data points in the proper category in the future. A hyperplane is the optimal choice boundary.



Model performance:

Model can be evaluated by various metrics such as:

1. **Confusion Matrix-**
   The confusion matrix is a table that summarizes how successful the classification modelis at predicting examples belonging to various classes. One axis of the confusion matrix is the label that the model predicted, and the other axis is the actual label.

2. **Precision/Recall**-
   Precision is the ratio of correct positive predictions to the overall number of positive predictions : TP/TP+FP

   Recall is the ratio of correct positive predictions to the overall number of positive examples in the set: TP/FN+TP

3. **Accuracy-**
   Accuracy is given by the number of correctly classified examples divided by the total number of classified examples. In terms of the confusion matrix, it is given by: TP+TN/TP+TN+FP+FN

Hyper parameter tuning:

Hyperparameters are sets of information that are used to control the way of learning an algorithm. Their definitions impact parameters of the models, seen as a way of learning, change from the new hyperparameters. This set of values affects performance, stability and interpretation of a model. Each algorithm requires a specific hyperparameters grid that can be adjusted according to the business problem. Hyperparameters alter the way a model learns to trigger this training algorithm after parameters to generate outputs.

We used Grid Search CV

Grid Search CV-Grid Search combines a selection of hyperparameters established by the scientist and runs through all of them to evaluate the model’s performance. Its advantage is that it is a simple technique that will go through all the programmed combinations. The biggest disadvantage is that it traverses a specific region of the parameter space and cannot understand which movement or which region of the space is important to optimize the model.

Conclusion:

That's it! We reached the end of our exercise.

Starting with loading the data so far we have done EDA, feature selection and then model building.

In all of these models our accuracy revolves in the range of 85 to 96%.

And there is no such improvement in accuracy score even after hyperparameter tuning.

So the accuracy of our best model is 96% for SVM model which really good for this dataset..