Mobile Price Range Prediction

Jaya Vishwakarma, Kavya Sharma,
Priyvrat Sharma, Richa Pandya
Data science trainees
Alma Better, Bangalore

ABSTRACT:

A predictive model was built from the mobile price range dataset to predict the range of a mobile phone based on different features. The model was built using different algorithms such as logistic regression, random forest, decision trees, XG boost and KNN. The performance of these models was closely checked to figure out the best model for the mobile price range prediction. Logistic regression seemed to predict the price range more closely than the other algorithms. XG boost seemed to over fit the data.

Keywords: *predictive, Logistic regression, random forest, decision trees, XG boost, KNN, overfit, best model*

INTRODUCTION:

Price is the most effective attribute of marketing and business. It is the most important factor that decides the sales of that product. Mobile technology is a technology where users go, this technology also goes. The mobile phone is stimulating one of the most important technological revolutions in human history. This portable technology consists of two-way communication, computing and networking technology. There are many features which are important to consider a mobile price like brand, display, resolution, ram, camera, processor, chipset etc. So, it becomes very important for a company to decide on which features it should focus to maximize the sales of that particular mobile phone in the market. The current project aims to figure out which of the attributes are the most important ones in predicting the price of the mobile phone in the market based on the data provided by Alma Better. The dataset contained a list of columns/features including total energy or battery can store in one time measured in MAh, presence of Bluetooth or not, speed at which microprocessor executes instructions, has dual sim support or not, front Camera mega pixels etc. The model was trained and validated using supervised ML models such as Naïve Bayes, Random forest, XG boost, KNN and Logistic Regression.

PROBLEM STATEMENT:

Develop a Supervised learning model using classification algorithms to predict the price range of mobile phones in the ranges:

0 - low cost

1 - medium cost

2 - high cost

3 - very high cost

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 DESCRIPTION:

The data description phase starts with an initial data collection and proceeds with activities in order to get familiar with the data. Identifying data quality problems, discovering first insights into the data and detecting interesting subsets to form hypotheses from hidden information are activities of this step. The data was taken from mobile Provider Company. It has 2000 rows and 21 columns. The columns that we have in our dataset are:

Battery_power- Total energy a battery stored 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- Support 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
Three_g- Support 3G or not
Touch_screen- Has touch screen or not
Wifi- Has wifi or not
Price_range- This is the target variable with values of 0(low cost), 1(medium cost),2(high cost) and 3(very high cost)

UNDERSTADING DATA

This is one of the most crucial steps that deals with descriptive statistics and analysis of the data. The main tasks involve summarizing the data, finding the hidden correlation and relationships among the data, developing predictive models, evaluating the models, and calculating the accuracies.

DATA CLEANING

It is very important to get rid of the irregularities and clean the data after sourcing it into our system.
Irregularities are of different types of data.

- Missing Values
- Incorrect Format
- Incorrect Headers
- Anomalies/Outliers

DATA TRANSFORMATION

Data transformation is the process of normalizing and aggregating the data to further improve the efficiency and accuracy of data mining.

DATA PREPROCESSING:

Dataset may contain noise, missing values and inconsistent data. Thus pre-processing of data is essential to improve the quality of data and time required in the data mining.

HANDLING OUTLIERS:

Outliers are data points that diverge from other observations for several reasons. During the EDA phase, one of our common tasks is to detect and filter these outliers. The main reason for this detection is that the presence of such outliers can cause serious issues in statistical analysis.

EDA

If we want to explain EDA in simple terms, it means trying to understand the given data much better, so that we can make some sense out of it. we are using univariate analysis to describe key characteristics of each feature including, minimum and maximum value, average, standard deviation and others. It was also used to produce a value distribution and identify missing values and outliers.

GRAPHICAL REPRESENTATION OF THE RESULTS: This step involves presenting the dataset with respect to the target feature in the form of graphs, summary tables, Bar chart, Scatter plot, Area plot, stacked plot Pie chart, Table chart, Polar chart, Histogram etc.

ALGORITHMS:

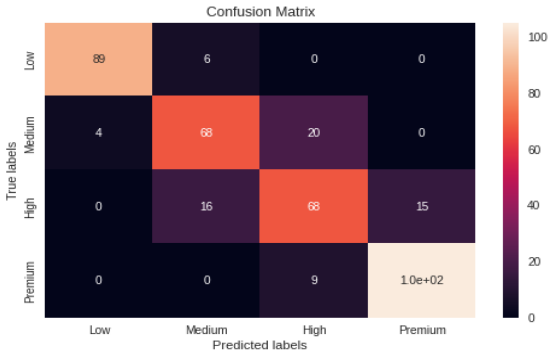
1. Naïve Bayes:

Naïve Bayes is a supervised machine learning model majorly used in solving classification problems. Supervised machine learning models are those where we use text classification that includes a high-dimensional training dataset.

Gaussian naive bayes: The Gaussian model assumes that features follow a normal distribution. This means if predictors take continuous values instead of discrete, then the model assumes that these values are sampled from the Gaussian distribution.

Score Metrics:

```
Evaluation metrices on the test data
Accuracy : 0.825
Recall : 0.825
Precision : 0.8239428786535121
F1 : 0.8242390777915095
[0.825, 0.825, 0.8239428786535121, 0.8242390777915095]
```

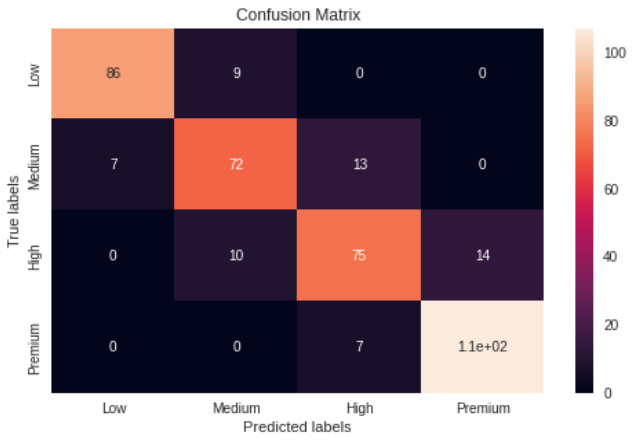


2. DECISION TREE:

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.

Score Metrics

```
Evaluation metrices on the test data
Accuracy : 0.84
Recall : 0.84
Precision : 0.8406531109445278
F1 : 0.8402461172404688
[0.84, 0.84, 0.8406531109445278, 0.8402461172404688]
```

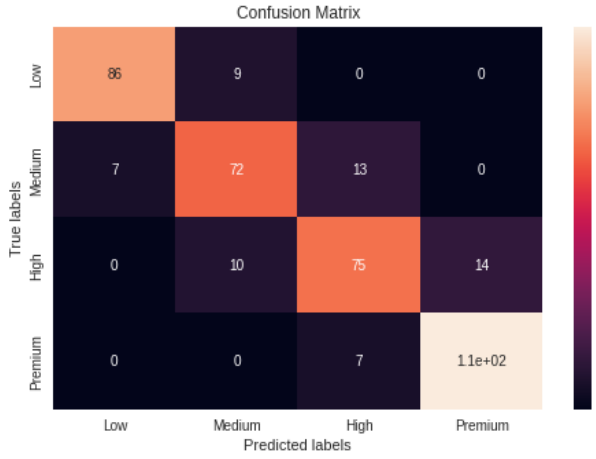


3. RANDOM FOREST:

Random Forest is a bagging type of Decision Tree Algorithm that creates a number of decision trees from a randomly selected subset of the training set, collects the labels from these subsets and then averages the final prediction depending on the most number of times a label has been predicted out of all.

Score Metrics

```
Evaluation metrices on the test data
Accuracy : 0.84
Recall : 0.84
Precision : 0.8406531109445278
F1 : 0.8402461172404688
[0.84, 0.84, 0.8406531109445278, 0.8402461172404688]
```



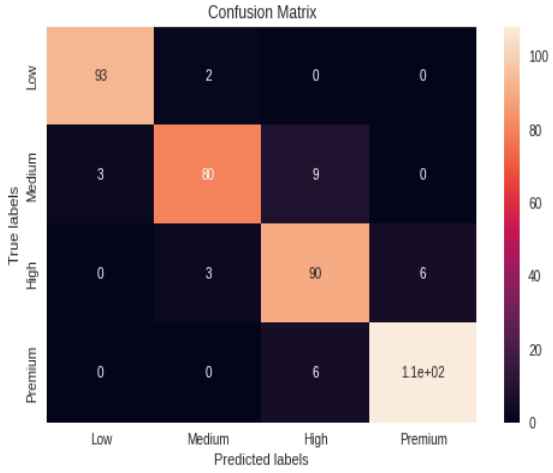
4. GRADIENT BOOSTING (XG Boost):

XG Boost stands for Extreme Gradient Boosting. The term gradient boosting consists of two sub-terms gradient and boosting. We already know that gradient boosting is a boosting technique. Let us see how the term 'gradient' is related here.

Gradient boosting re-defines boosting as a numerical optimisation problem where the objective is to minimise the loss function of the model by adding weak learners using gradient descent. Gradient descent is a first-order iterative optimisation algorithm for finding a local minimum of a differentiable function. As gradient boosting is based on minimising a loss function, different types of loss functions can be used resulting in a flexible technique that can be applied to regression, multi-class classification etc

Score Metrics

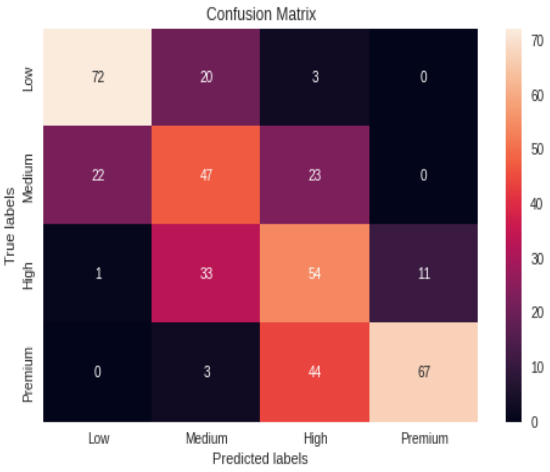
```
Evaluation metrices on the test data
Accuracy : 0.93
Recall : 0.93
Precision : 0.9300236392688487
F1 : 0.9299368559017597
[0.93, 0.93, 0.9300236392688487, 0.9299368559017597]
```



5. KNN: KNN is a method which is used for classifying objects based on closest training examples in the feature space. KNN is the most basic type of instance-based learning or lazy learning.

Score Metric

```
Evaluation metrices on the test data
Accuracy : 0.6
Recall : 0.6
Precision : 0.637541406682888
F1 : 0.6096435157238129
[0.6, 0.6, 0.637541406682888, 0.6096435157238129]
```

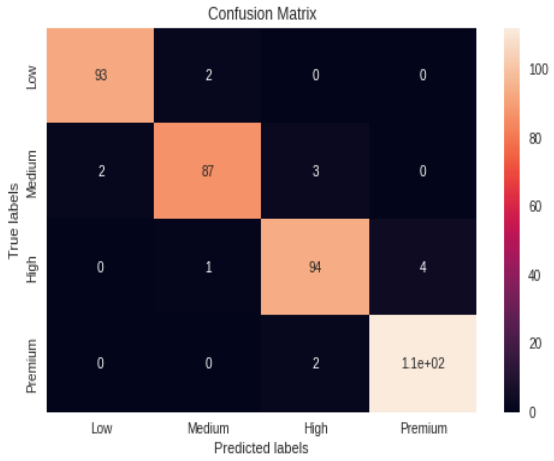


6. LOGISTIC REGRESSION

Logistic regression estimates the probability of an event occurring, such as voted or didn't vote, based on a given dataset of independent variables. Since the outcome is a probability, the dependent variable is bounded between 0 and 1.

Score Metric

```
Evaluation metrices on the test data
Accuracy : 0.965
Recall : 0.965
Precision : 0.9650057471264368
F1 : 0.9649553272814143
[0.965, 0.965, 0.9650057471264368, 0.9649553272814143]
```



CONCLUSIONS:

Different features like Battery power, clock speed, dual sim, mobile depth, mobile weight, pixel height, pixel width, ram, secondary camera, talk time got linear relationship with our dependent variable price range. It was seen that ram has the highest impact on the price of the mobile. Surprisingly, Logistic regression performed well in this classification problem. It was also found that there's some over fitting in case of XG Boost. Also, Random Forest and Decision Tree models were tested which performed well as compared to the other models like KNN and Naive Bayes.

Hence, it was concluded that the Logistic model was the best among all.

REFERENCES:

- Alma Better Recorded Classes
- Stack overflow