Mobile Range Classification

Logistic Regression

# Domain

Telecom, Mobile

# Business Context

Market size of mobile phones is growing everyday so is the competition. Hence, companies try to maximize the sales which depends on various factors like demand, technology, marketing, brand, availability, user experience, service, price, etc. Electronics companies also focus on improving their products in order to capture the maximum market size. As we understand that selling price makes a huge difference when it comes to sales and profits.

However, estimating an optimal price for a new mobile phone can be a tricky task especially when you are new in the business or when you want to launch a new kind of mobile phone in the market. So, businesses utilize their domain knowledge and historical information to classify an optimal price range for a product.

Objective

To classify the price range for a mobile phone based on sales data collected from various similar electronics companies. To find out the relationship between features of the mobile phone(eg:- RAM, Internal Memory, etc) and its selling price.

Here, the data collected is from various electronic firms. Dataset includes information about mobile features like - battery, RAM, Mobile\_wt, 4g, etc. Target column has four labels of price range. We will train a logistic regression algorithm to build a model that can predict the price range of a new mobile phone.

Dataset description

Dataset has 21 columns and 2000 rows

* Id : ID
* 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).

Steps

1. You are given data of 2000 mobile phones with the information of 21 features. And you are supposed to analyze the data and answer the following questions.
   1. Check a few observations and mention names of the categorical features
   2. What is the percentage of each price range in the available train? Is the data balanced?
   3. Check for missing values. Impute the missing values if there is any
   4. You are told that the ‘ram’ of a mobile influences its price the most. Can you prove or disprove this claim?
   5. List out the features which do not have linear relationship with price range
   6. Check if there are any outliers in the dataset and mention your comments
2. You are asked to build a Logistic Regression model to predict the price range. Kindly follow the below steps and answer the following -
   1. Segregate the dependent column from the data frame
   2. Split the dataset into training and testing set ( 70:30 split)
   3. Train logistic regression on training set and predict price range for the test data
   4. Check accuracy score and display confusion matrix. Do you think the logistic model is good to move into production?
   5. Is the model overfit or underfit?
3. From the above step you conclude that the model is not good for production. What all steps can be taken to improve the accuracy?
   1. Standardize the data and check the accuracy scores and confusion matrix
   2. Does the accuracy improve?
4. What are your findings from the exercise?