Mobile Phones Selling Price Report

Proposal for final project (MDSA Winter 2023)

Yongpeng Fu (10182778), Rudy Brown (), Jose Palacios (),
Stuart Finley(), Andrii Voitkiv()

2023-01-29

Contents

Chapter 1: Introduction

Mobile phones are everywhere, so are the prices. Despite still having the word "phone" in the name, a typical modern smartphone has much more features than just to make and receive calls. They are boasting a staggering range of features, like brand, memory, storage, camera, resolution, just to name a few. And as you can imagine, with all this new technology and features jam packed in one little device costs money, costs a lot of money. A 2020 review of premium mobile phones shows a staggering *490%* rise in the last two decades.

With so many mobile phones on the market, it can be difficult to decide which one you want to buy. As a customer, we are particularly interested in finding some relation between all these features and its selling price. To this purpose, we collected the MobilePhone's dataset from Kaggle and apply a set of statistical analysis hoping to answer some guiding questions:

1. Can we estimate the true average price for mobile phones?
2. What is the impact of each mobile phone's feature on the selling price?
3. Can a classification model to distinguish the selling price range?
4. Can we build a decent model to predict the selling price for a mobile phone?

Chapter 2: Dataset and Scope of Analysis.

~~We chose a dataset weighting by *simplicity*. That is, we would like to maximize the learning experience applying class content to a toy/stylized modelthat may or may not have any practical use.~~

Dataset

The dataset consists of 8 columns and 28,036 rows and no missing values. These 8 columns are:

- **Model**: categorical variables with sub-classes. These names include the color of the unit and its storage capacity. The latter being also listed as a separate column. - **Independent Variable**
- **Company**: categorical variable. Name of the phone's manufacturer. - **Independent Variable**
- **Price**: continuous variable. Units in Indian Rupees. - **Dependent Variable**
- **Rating**: continuous variable. Units in Indian Rupees. - **Independent Variable**
- **Number of ratings**: discrete variable: a simple count. - **Independent Variable**
- **Total reviews**: discrete variable: a simple count. - **Independent Variable**
- **RAM size**: categorical variable. RAM specification of the phone. - **Independent Variable**
- **ROM size**: categorical variable. Storage (non-volatile memory) capacity of the phone. - **Independent Variable**

The data was loaded into RStudio using the "read.csv" function. The data we needed for our analysis include the dependent variable "Price" (in Indian Rupee) and a list of independent variables listed above. The first few records of the dataset is as follows:

```
mobile_dataset <- read.csv("./Updated_Mobile_Dataset.csv")
mobile_dataset %>% head(4)
```

```
##                                     Model  Company Price Rating No_of_ratings
## 1 Infinix HOT 20 Play (Luna Blue, 64 GB)  Infinix  8199    4.3           505
## 2      MOTOROLA e40 (Carbon Gray, 64 GB) MOTOROLA  7999    4.1         56085
## 3        MOTOROLA e40 (Pink Clay, 64 GB) MOTOROLA  7999    4.1         56085
## 4          POCO C31 (Shadow Gray, 64 GB)     POCO  7499    4.3        183688
##   TotalReviwes RamSize RomSize
## 1           52    4 GB   64 GB
## 2         5600    4 GB   64 GB
## 3         5600    4 GB   64 GB
## 4        11185    4 GB   64 GB
```

As the first step of investigation, we did the following work to clean up the dataset.

1. Remove any duplicates in the dataset;
2. Because **Model** column contains sub-class of a mobile phone, we decide to further break it down to *Model* and *Color*;
3. Add additional column to segment the **Price** into 4 different levels.

```
import pandas as pd
data = pd.read_csv('./Updated_Mobile_Dataset.csv')

def map_to_cat(price):
    if price < 7500:
        return 'Low'
    elif 7500 <= price < 15000:
        return 'Medium'
    elif 15000 <= price < 30000:
        return 'High'
    elif price >= 30000:
        return 'Very high'

aug_model = data['Model'].copy()
aug_color = [None]*len(aug_model)
aug_price_category = [None]*len(aug_model)
for i,r in data.iterrows():
    aug_model[i] = str(aug_model[i]).replace(r['Company'], '').strip()
    aug_price_category[i] = map_to_cat(r['Price'])
    if (r['Model'].find(',') != -1) and (r['Model'].find('(') != -1):
        aug_color[i] = aug_model[i].split('(')[1].split(',')[0]
        aug_model[i] = aug_model[i].split('(')[0]

data['aug_model'] = aug_model
data['aug_color'] = aug_color
data['aug_price'] = aug_price_category

data = data.drop_duplicates().reset_index(drop=True)
data = data[data['Company'] != 'Nothing']
data['Company'] = data['Company'].str.capitalize()

data.to_csv('./augmented_dataset.csv', index=False)
data
```

```
##                                        Model   Company  ...     aug_color  aug_price
## 0    Infinix HOT 20 Play (Luna Blue, 64 GB)   Infinix  ...     Luna Blue     Medium
## 1         MOTOROLA e40 (Carbon Gray, 64 GB)  Motorola  ...   Carbon Gray     Medium
## 2           MOTOROLA e40 (Pink Clay, 64 GB)  Motorola  ...     Pink Clay     Medium
## 3             POCO C31 (Shadow Gray, 64 GB)      Poco  ...   Shadow Gray        Low
## 4        MOTOROLA G32 (Mineral Gray, 64 GB)  Motorola  ...  Mineral Gray     Medium
## ..                                      ...       ...  ...           ...        ...
## 735                            Kechaoda K16  Kechaoda  ...          None        Low
## 736              LAVA Z2 (Flame Red, 32 GB)      Lava  ...     Flame Red     Medium
## 737                POCO F1 (Rosso Red, 64 GB)      Poco  ...     Rosso Red        High
## 738            OPPO A54 (Starry Blue, 128 GB)      Oppo  ...   Starry Blue        High
## 739                              Kechaoda K200  Kechaoda  ...          None         Low
##
## [736 rows x 11 columns]
```

```
mobile_dataset <- read.csv("./Updated_Mobile_Dataset.csv")
cat("Break down Model column into Model only and Color columns:\n")
```

```
## Break down Model column into Model only and Color columns:
```

```
#Step 1: remove the Company name from Model
Model_no_Company <- stringr::str_remove(mobile_dataset$Model,mobile_dataset$Company)
↪  %>% trimws(., which = c("both"))
#Step 2: remove anything after parentheses to get Model only info
mobile_dataset$Model_Only <- stringr::str_replace(Model_no_Company, "
↪  \\s*\\([^\\)]+\\)", "")
#Step 3: Get the color information from inside the last ()
#Step 3.1: Get the parenthesis and what is inside from the last ()
Model_no_Company_parenthesis <- stringr::str_extract(Model_no_Company,
↪  "(?<=\\()([^()]*?)(?=\\)[^()]*$)")
#step 3.2: Get the color by just retaining the info before ,
mobile_dataset$Color <- gsub(",.*$", "", Model_no_Company_parenthesis)
#Step 4: Remove duplicated rows in the dataset
mobile_dataset <- mobile_dataset[!duplicated(mobile_dataset),]
#Step 5: cut the price based on the percentile into 4 different levels
mobile_dataset <- mobile_dataset %>% mutate(Price_Level = ntile(Price, n = 4))
#Step 5.1: map each number level to the character
from <- c(1,2,3,4)
to <- c("Low", "Medium", "High", "Very High")
mobile_dataset$Price_Level <- mapvalues(mobile_dataset$Price_Level, from = from, to
↪  = to)
```

After cleaning and breaking down columns, the dataset now consists of 11 columns and 740 rows and no missing values. These 11 columns are:

- **Model**: categorical variables with sub-classes. These names include the color of the unit and its storage capacity. The latter being also listed as a separate column. - **Independent Variable**
- **Company**: categorical variable. Name of the phone's manufacturer. - **Independent Variable**
- **Price**: continuous variable. Units in Indian Rupees. - **Dependent Variable**
- **Rating**: continuous variable. Units in Indian Rupees. - **Independent Variable**
- **Number of ratings**: discrete variable: a simple count. - **Independent Variable**
- **Total reviews**: discrete variable: a simple count. - **Independent Variable**
- **RAM size**: categorical variable. RAM specification of the phone. - **Independent Variable**
- **ROM size**: categorical variable. Storage (non-volatile memory) capacity of the phone. - **Independent Variable**
- **aug_model**: categorical variable. It only contains the Model information for a mobile phone. - **Independent Variable**
- **aug_color**: categorical variable. The color of a mobile phone. - **Independent Variable**
- **aug_price**: categorical variable. The price level of a mobile phone, with levels of "Low", "Medium", "High", "Very High" - **Independent Variable**

The data was again loaded into RStudio using the "read.csv" function. The first few records of the dataset is as follows:

```
mobile_dataset <- as_tibble(mobile_dataset)
cat("The dimension of the dataset:\n")
```

```
## The dimension of the dataset:
```

```
dim(mobile_dataset)
```

```
## [1] 740  11
```

```
mobile_dataset %>% head(4)
```

```
## # A tibble: 4 x 11
##   Model           Company Price Rating No_of_ratings TotalReviwes RamSize RomSize
##   <chr>           <chr>   <int>  <dbl>         <int>        <int> <chr>   <chr>
## 1 Infinix HOT 2~ Infinix  8199    4.3           505           52 4 GB    64 GB
## 2 MOTOROLA e40 ~ MOTORO~  7999    4.1         56085         5600 4 GB    64 GB
## 3 MOTOROLA e40 ~ MOTORO~  7999    4.1         56085         5600 4 GB    64 GB
## 4 POCO C31 (Sha~ POCO     7499    4.3        183688        11185 4 GB    64 GB
## # ... with 3 more variables: Model_Only <chr>, Color <chr>, Price_Level <chr>
```

Scope of Analysis

TODO: Hash out modelling. I recommend a diagram using a tool like lucidcharts.com. See the templates section. For instance (see Figure 1 below).

The dataset and detailed analysis can be found at this repository.

Chapter 5: References

TODO

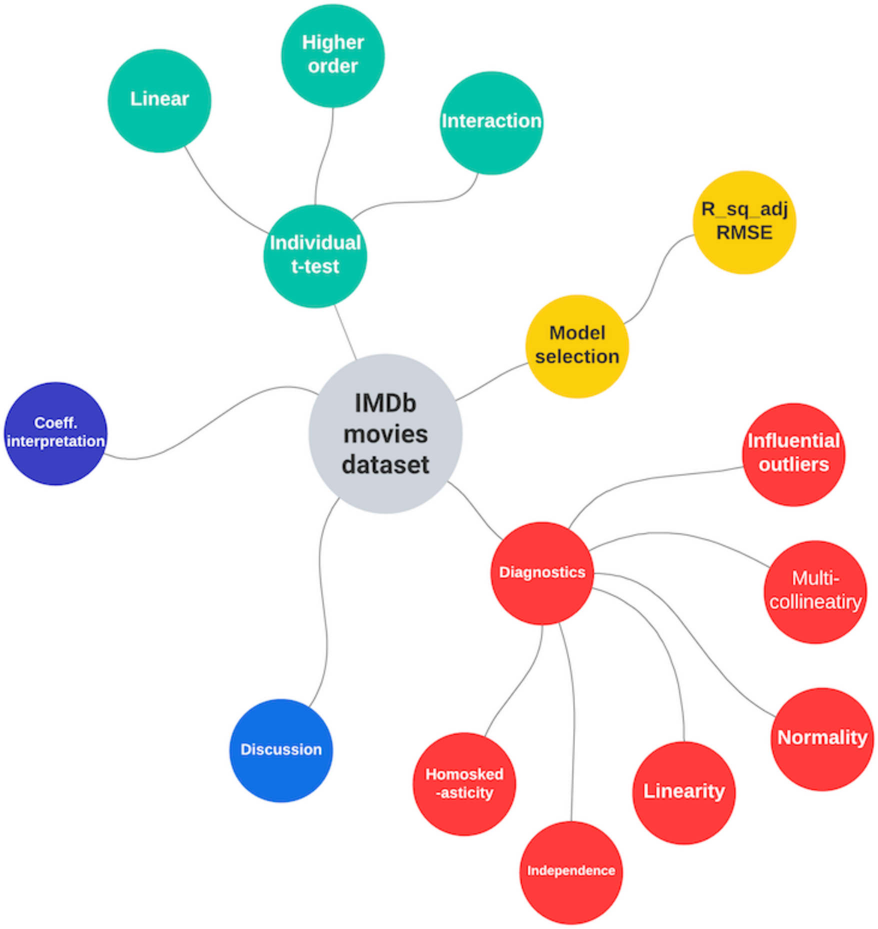

Figure 1: Example diagram