Smartphone Price Prediction

GitHub Link

Group No: 3
Naimul Haque 14.02.04.080
Samin Shahriar Tokey 14.02.04.066

February 08, 2018

1 Introduction and Background

Pricing is one of the building blocks of marketing that appears to be easy to understand, but is probably one of the most difficult. Many think it is easy because we all buy products that have prices, and many believe all you have to do is sell the product for more than it costs you to earn a profit.

This report aims at explaining how to use different machine learning models that can predict the price of different smartphones using **'smartphone dataset'**.

1.1 Predictive modeling

Predictive modeling is the problem of developing a model using data-set to make a prediction on new data where we do not have the answer.

Predictive modeling can be described as the mathematical problem of approximating a mapping function (f) from input variables (X) to output variables (y). This is called the problem of function approximation.

1.2 Regression Predictive Modeling

Regression predictive modeling is the task of approximating a mapping function (f) from input variables (X) to a continuous output variable (y).

In our case,our models take the features of smartphones as input variables and predict the prices as the output variable.

1.3 Machine Learning Models

Regression is concerned with modeling the relationship between variables that is iteratively refined using a measure of error in the predictions made by the model.

Among the very popular machine learning models,we used following classifiers:

1. **Random Forest Classifier**
2. **Bagging Classifier**
3. **Decision Tree Classifier**
4. **Linear Regression**
5. **Logistic Regression**
6. **Lasso**

2 Smartphone Dataset

2.1 Abstract

The data are collected from the internet,which concerns the different quality score of 103 different devices.

Data Set Characteristics:	Multivariate	**Number of Instances:**	103
Attribute Characteristics:	Categorical, Real	**Number of Attributes:**	4
Associated Tasks:	Regression	**Missing Values?**	No

Table 1: Smartphone dataset

2.2 Data Set Information:

Each records are searched in the Google and datas are carefully selected from different websites.

2.3 Attribute Information:

1. **Geekbench Single Core Score** :

   The Single Thread CPU benchmark, like all processor benchmarks attempts to estimate how quickly a processor is able to perform a wide variety of calculations. The test issues as series of complex instructions to the processor and times how long the processor takes to complete the tasks. The faster the processor is able to complete the tasks, the higher the benchmark score. The GeekBench Single Thread CPU test only runs one stream of instructions rather than multiple parallel streams per core. The majority of consumer applications (MS World, Internet Explorer, Google Chrome and most games), although multi threaded, rarely utilize more than one thread at a time, so this test, like any single threaded benchmark, can be seen as a reasonable real world test for typical consumer workloads.

2. Geekbench Multi Core Score :

   This estimates the overall performance of all the cores working together and gives a score based on the result.

3. **DxO Mark Camera Rating** :

   DxOMark is a website providing image quality ratings for standalone cameras, lenses, and mobile devices that include cameras. It is owned by DxOMark Image Labs. As smartphones overtake point-and-shoot cameras, DxO Labs started testing smartphones and other mobile devices in 2011 and introduced DxOMark Mobile in 2012. A major update was made in September 2017, adding tests designed to stress the capabilities of current-model smartphones such as lower-light shooting, telephoto zoom, depth effect, and bokeh.

   DxOMark Mobile Overall Score is the headline number reported for each tested device, and consists of: DxOMark Mobile Photo and DxOMark Mobile Video.

   DxOMark's Mobile Photo score is composed of these Categories:

   - Exposure and contrast
   - Color
   - Autofocus
   - Texture
   - Noise
   - Artifacts
   - Flash
   - Zoom
   - Bokeh

   DxOMark's Mobile Video score includes six of the same Sub-scores as DxOMark's Mobile Photo score (Exposure, Color, Autofocus, Texture, Noise, and Artifacts), along with Stabilization.

   DxOMark's tests are conducted under a variety of lighting conditions, ranging from low light 5 lux to bright daylight outdoors.

4. **Release Date** :

   This attribute is the date of when the SmartPhone was released.

5. **Price** :

   Price of the device in BDT.

3 Project Details

3.1 Platform

The project is carried out in **python 3.0** environment using **Jupyter Notebook** which is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text.

3.2 Project Files

The project Folder following files:

1. **main.ipynb**
2. **dataset.csv**
3. **main.py**
4. **report.pdf**
5. **link.txt**

The **main.ipynb** file is a notebook document which can be opened by Jupyter Notebook. To execute the code segments,just keep pressing [Shift] and [Enter].

The **dataset.csv** contains the dataset to train the models.

The **main.py** can be executed by clicking on the file. The program will wait until you press any key.

The **AI**$_{p}$*roject.pdfisthereportdocumentofourprojectwrittenusinglatex. The***link.txt***containstheprojectandlatexviewforthereport.*

3.3 Tools and dependencies

We used python programming in implementing our project. There are a number of reasons we choose Python programming language.

One of the most commonly cited reasons is the syntax of Python, which has been described as both "elegant" and also "math-like."

Other is that there are many python libraries for data visualization and manipulation. There are also very useful libraries for doing linear algebra and other mathematical areas.

The libraries that we used are :

1. **numpy**

2. **sklearn**
3. **pandas**
4. **matplotlib**

The numpy is used for linear algebra.

The sklearn is used to implement different classifiers and K-folds.

The pandas is used for data manipulation.

The matplotlib is used for plotting graphs and data visualization.

4 Test and Train

4.1 Cross Validation

Cross Validation is a very useful technique for assessing the performance of machine learning models. It helps in knowing how the machine learning model would generalize to an independent data set. You want to use this technique to estimate how accurate the predictions your model will give in practice.

4.2 K-Fold Cross Validation

K-Fold Cross Validation is a common type of cross validation that is widely used in machine learning.

K-fold cross validation is performed as per the following steps:

Partition the original training data set into k equal subsets. Each subset is called a fold. Let the folds be named as f1, f2, . . . , fk . For i = 1 to i = k
Keep the fold fi as Validation set and keep all the remaining k-1 folds in the Cross validation training set.

We used 5-Fold Cross Validation in our project.

5 Performance Evalation

5.1 Coefficient of determination or R-squared

R-squared is a statistical measure of how close the data are to the fitted regression line.

The definition of R-squared is fairly straight-forward; it is the percentage of the response variable variation that is explained by a linear model. Or:

R-squared = Explained variation / Total variation

R-squared is always between 0 and 1:

- 0 indicates that the model explains none of the variability of the response data around its mean.
- 1 indicates that the model explains all the variability of the response data around its mean.

5.2 Result

We measured the performance of different classifier based on their R-squared score. The odder of classifiers based on their score:

1. Random Forest Classifier
2. Decision Tree Classifier
3. Bagging Classifier
4. Lasso
5. Linear Regression
6. Logistic Regression

6 Conclusion

The classifiers are expected to give better result with greater number of instances of the dataset. Hopefully,we look forward to collect more data and provide system with greater performace to use in practice.

References

[1] DxO Mark camera rating. `https://www.dxomark.com`. Last Accessed: 10-02-2018.

[2] GeekBench benchmark score. `https://www.geekbench.com/`. Last Accessed: 10-02-2018.

[3] **Google search engine.** `http://www.google.com//`. **Last Accessed: 10-02-2018.**

[4] **MuthoPhone price in bdt.** `http://www.muthofon.com//`. **Last Accessed: 10-02-2018.**

(1) (2) (4) (3)