**Smartphone Price Prediction**

# Group 7

Akhil Vaitla Janardhan 50418467

Arunkumar Muthuswamy 50430427

Nabeel Khan 50419151

Venkata Satya Surya Sai Vineet Atyam 50419767

Zhijie Xu 50222635

# **Abstract**

The objective of this study is to justify if the price of a smartphone with given features is reasonable, by utilizing raw data mined from *GSMArena.com*. To get significant features with the least amount of computational cost, different feature selection techniques are used. To achieve lowest error various regression approaches are applied and the models with lowest **MdAPE** (Median Absolute Percent Error) are compared with each other. The best feature selection algorithm and optimal regression for the given dataset are used to reach a conclusion. This activity can be applied to any sort of marketing and business to discover the best product (with minimum cost and maximum features). Future study should be done to expand on this work and develop a more sophisticated solution to the problem as well as a more precise tool for pricing estimation.

# **1. Introduction**

The most important marketing and commercial attribute for a consumer electronic product is its *price*. The primary goal of the project is to estimate cost of purchasing a smartphone. This study is merely the first step in the direction of the goal.

Statistics is a discipline of mathematics that provides tools for understanding the relationship between various properties of a data collection and identifying the uncertainty of a given feature based on previous data. MATLAB, Python, R, and more technologies are available for statistical learning. There are a variety of feature selection algorithms that may be used to select only the best features and reduce the size of the dataset. The problem's computational complexity will be reduced because of this. Nowadays, the smartphone is one of the most popular gadgets. Every day, new smart phones are released, each with a new version and more functionalities. Every day, hundreds of thousands of smart phones are purchased. As a result, the smartphone price prediction is a case study for the specified problem type, namely, finding the best product. The similar process can be employed to determine the cost of any product in any industry.

Many factors must be considered when estimating the price of a smartphone. Face recognition, fingerprint accessibility, battery capacity, mobile's size, thickness, internal memory, camera pixels, video quality, 5G support, and other features, for example, may play a role in determining the smartphone's pricing. As a result, we'll utilize a combination of the above characteristics to determine if the smartphone is reasonable or overpriced.

# **2. Methodology**

## **2.1 Data Mining**

Data was collected from GSMArena website for the analysis. Five different smartphone manufactures were selected based on relevance and corresponding data was collected from the website. Manual collecting the data was the simple solution in the initial phase but it turned out to be hard as the data collection was progressing because each smartphone page has to be visited and the features has to be extracted and finally all the data has to be combined into a single data source for the analysis to be initiated.

A better approach was to crawl through the websites and store the data and once data for different smartphones were collected then it can be aggregated. For crawling we used Python and built a script that will go through each smartphone page of the provided manufacturer and saves the collected data. Since GSMArena didn’t have any user-friendly API we ran into different problems and in order to avoid it we had to run the scripts with sleep in between to avoid DOS attacks.

At the end of the process, we had a single data source that will list down the features of the five smartphone manufactures.

Graphical user interface, text

Description automatically generated

Figure 1: Collected Raw Data (Only three features included in the figure. The figure only shows three features. In fact, there are 31 traits that combine textual and numerical representation.)

## **2.2 Data Processing**

After obtaining the dataset from GSMArena, data processing had to be done so that the analysis could be performed since data available would not be compatible with the formats required by the models in R. Major challenge in transforming the data involved handling the textual representation of the data, missing data, outliers like not all phones had state-of-the-art features, etc. For converting the textual data, two approaches were followed. One of them was to convert the value in the column into numeric and the second one was making sense out of that feature. For example, features like price were available in different formats like USD, Euro, etc. Different parsers were created in Excel, Python, R to extract the features from the columns. In certain cases, deriving features was also possible like from CPU spec features like the number of cores, clock speed, etc were extracted. In case of missing data, median values were used to replace them, so that other features of the observation can be effectively used. In the end, there were 31 features that were cleaned and processed for the next phase which is the analysis.

Table

Description automatically generated

Figure 2: Pre-processed Dataset

### *2.2.1 Data Transformation*

We mined raw data from GSM arena website which resulted in raw data in text format. We parsed features like Battery Size, RAM, Internal Memory etc

To utilize CPU and GPU we made a feature called CPU score, and GPU score from **Centurion Mark** which is one of the industry-leading benchmarking techniques to evaluate the performance of a processor.

We used fuzzy logic to map existing CPU and GPU name to the closest GPU name and the corresponding Centurian Mark (score) as illustrated in the table below.

Graphical user interface

Description automatically generated with low confidence

### **Processed Data**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Feature** | **min** | **mean** | **median** | **max** |
| **Thickness (mm)** | 5.20 | 8.86 | 8.60 | 73.50 |
| **NFC** | 0.00 | 0.54 | 1.00 | 1.00 |
| **Battery (mAh)** | 1500.00 | 4151.99 | 4200.00 | 12000.00 |
| **Bluetooth** | 4.00 | 4.78 | 5.00 | 5.20 |
| **PPI** | 127.00 | 363.53 | 395.00 | 577.00 |
| **Price (USD)** | 40.00 | 353.37 | 245.00 | 4681.60 |
| **Announced** | 2015.92 | 2019.80 | 2020.17 | 2021.92 |
| **#5g\_bands** | 0.00 | 1.63 | 0.00 | 14.00 |
| **#4g\_bands** | 1.00 | 12.41 | 11.00 | 27.00 |
| **#3g\_bands** | 2.00 | 4.46 | 5.00 | 7.00 |
| **#2g\_bands** | 1.00 | 3.95 | 4.00 | 4.00 |
| **Total\_bands** | 7.00 | 22.45 | 20.00 | 47.00 |
| **#cameras** | 0.00 | 2.32 | 2.00 | 5.00 |
| **nanometers** | 5.00 | 12.64 | 11.00 | 28.00 |
| **Weight** | 126.00 | 188.76 | 185.00 | 2230.00 |
| **Single-MP** | 2.00 | 13.58 | 10.00 | 40.00 |
| **Size** | 3.80 | 6.24 | 6.40 | 17.30 |
| **Screen to body ratio** | 56.00 | 80.30 | 82.30 | 95.90 |
| **Internal** | 8.00 | 83.28 | 64.00 | 512.00 |
| **Ram** | 1.00 | 4.59 | 4.00 | 12.00 |
| **Wattage** | 0.00 | 22.43 | 18.00 | 100.00 |
| **Wireless Charging** | 0.00 | 0.00 | 0.00 | 1.00 |
| **CPU\_score** | 42.00 | 94.84 | 98.00 | 158.00 |
| **GPU\_score** | 27.40 | 82.81 | 81.60 | 131.30 |

Table 1: Summary of feature variables

# **3. Analysis & Results**

## **3.1 Analysis**

### *3.1.1 Correlation Heatmap*

We observed that the average correlation between features was ~0.26, and in the general features were not very highly correlated with each other. Feature reduction was not needed since we had low correlation between features.

A picture containing text, scoreboard, colorful

Description automatically generated

Figure 4: Correlations between the different attributes using heat map

### *3.1.2 Price Distribution*

Our dataset consists of Android phones. Since Android market is heavily dominated by relatively inexpensive phones, our dataset consists of mainly smartphones under 600$.



Table 2: Price distribution

Chart, histogram

Description automatically generated

### Figure 5: Price distribution chart

### *3.1.3 5G Adoption*

We have observed a clear trend in the 5G adoption. 5G has been pushed to the market aggressively and almost 50% of phones in 2021 have 5G capability.

Chart, pie chart

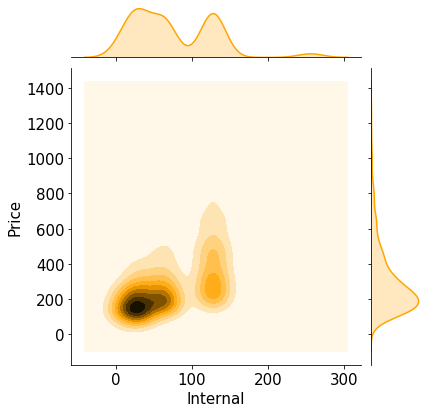
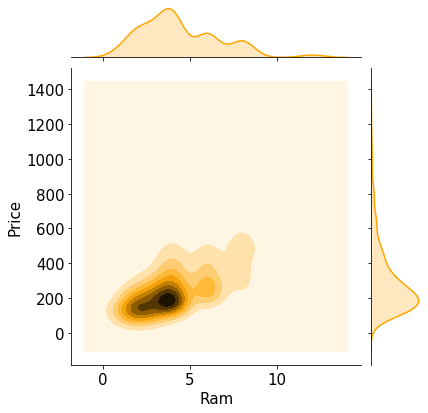
Description automatically generated Chart, line chart

Description automatically generatedFigure 6: Pie chart depicting % of smartphones Figure 7: Line plot of smartphone V/S 5G availability

### *3.1.4 RAM & Internal Memory vs Price*

Although it seems that the most important features related to price are RAM and Internal Memory, we did not find these as top feature in feature importance

We observed a positive correlation between RAM and Price, and Internal Memory and Price.



### Figure 8: Ram and Internal Storage V/S Price

### *3.1.5 CPU & GPU Score vs Price*

We created CPU score feature from CPU details, and we found that CPU score was most important feature when predicting the price of a smartphone.

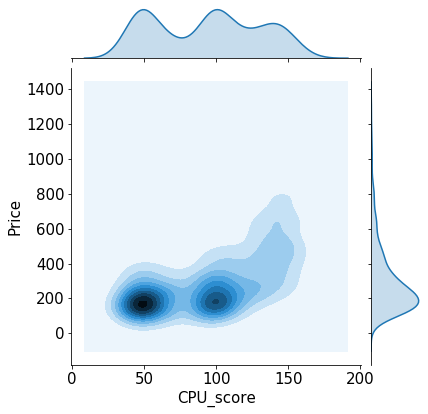
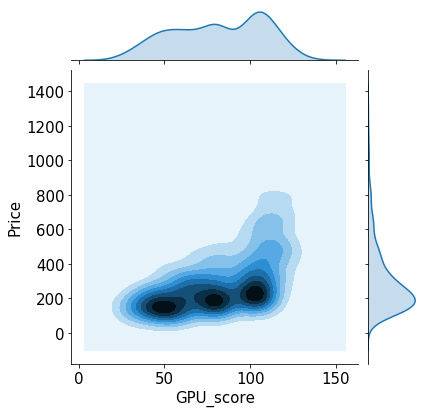


Figure 9: GPU score and CPU score V/S Price

# **4. Conclusion**

## **4.1 Results**

**Text

Description automatically generatedMedian Absolute Percent Error (MdAPE)**, also known as median absolute percentage deviation (MdAPD), is a measure of prediction accuracy of a forecasting method in statistics. It usually expresses the accuracy as a ratio defined by the formula: Mean absolute percentage error is commonly used as a loss function for regression problems and in model evaluation, because of its very intuitive interpretation in terms of relative error.

MdAPE = Median( )

where *At* is the actual value and *Ft* is the forecast value. Their difference is divided by the actual value *At.*

We tried different models and we were able to achieve best MdAPE as **16.25%** with **Random Forest**



## **4.2 Future Work Extension**

Artificial intelligence techniques can be used to improve accuracy and accurately estimate product prices. Software or mobile app that predicts the market price of a newly announced product can be produced.

More and more cases should be added to the data set to obtain maximal accuracy and predict more accurately. Additionally, picking more appropriate features can improve accuracy. To get higher accuracy, the data collection should be huge, and more appropriate features should be used.

To achieve maximum accuracy and predict more accurate, more and more instances should be added to the data set. And selecting more appropriate features can also increase the accuracy. So, data set should be larger and more appropriate features should be selected to achieve higher accuracy

# **5. References**

* *GSMArena.com*. (n.d.). Mobile Phone Reviews, News, Specifications and More... Retrieved December 10, 2021, from *https://www.gsmarena.com/*
* Vikramb. (2018, June 10). Mobile price prediction. *Kaggle*. *https://www.kaggle.com/vikramb/mobile-price-prediction*
* *Mobile GPU rankings 2021 (adreno/mali/powervr)*. (n.d.). Tech Centurion. Retrieved December 10, 2021, from *https://www.techcenturion.com/mobile-gpu-rankings*
* Schott, M. (2020, February 27). Random forest algorithm for machine learning - Capital one tech - Medium. *Capital One Tech*. *https://medium.com/capital-one-tech/random-forest-algorithm-formachine-learning-c4b2c8cc9feb*
* Alpaydin, E. (2014b). *Introduction to machine learning*. MIT Press. B.Balakumar,
* P.Raviraj, & V.Gowsalya. (n.d.-a). *Mobile Price prediction using Machine Learning Techniques*. Retrieved December 10, 2021, from *https://scholar.google.co.in/citations?view\_op=view\_citation&hl=en&user=CH8oJbsAAAAJ&citation\_for\_view=CH8oJbsAAAAJ:LkGwnXOMwfcC*