New Phone, Who's This?

An analysis of cell phone features and their relationship with pricing.

Abstract

Ever since their creation in 1973, cellphones have evolved from a concrete block, used specifically to call others, to small computers which manage our everyday activities. We spend about 5.4 hours a day navigating our devices.[1] Due to the advancement of technology and increased popularity, mobile phones have skyrocketed in price. Finding the best value for a phone can be difficult; thus, understanding the relationship between a phone's characteristics and cost is beneficial to both consumers and manufacturers.

Many statistical analyses have been performed on mobile phone pricing; however, they focus on regression modeling[2], phone usage[3], or differences between competitors[4]. In my study, I analyzed cell phone characteristics. I developed a random forest model that classified a phone into one of four price ranges (Low, Below Average, Above Average and High) with 85% accuracy. Also, we discovered that a phone's amount of RAM storage has an overwhelming amount of influence on the phone's pricing.

Introduction

Look to your left. Now, look to your right. Lastly, reach into both of your pockets. You identified your phone following one of those instructions, right? The development of technology has increased cell phone popularity over the past 30 years. At first, mobile phones were developed to allow for communication wherever needed. Now, devices allow users to video chat, call, play games independently or simultaneously. Users can develop careers, maintain relationships, or store memories with a touch of their screen. For example, Lil Nas X launched his music career when his Grammy award winning song 'Old Town Road' rose to fame on the mobile app Tiktok[5]. The increase in productivity makes cell phones vital for every person.

Because cell phones are now used for many of our activities, our devices require quality hardware; therefore, technology costs increase for both the consumer and the manufacturer. From 2015 to 2020, according to Eric Zeman, phone prices have doubled in price[5]. The price not only reflects manufacturing costs, but also the decrease in consumer demand. Improved quality over the past 5 years has increased inelasticity for cell phone purchases. Therefore, manufacturers tend to reflect this trend in their pricing. Previous analyses have conducted similar studies. However, many were performed within a limited scope. Either inflation wasn't accounted for, a single machine learning technique was tested, or the data was outdated. Our study not only compares characteristics of different mobile phones but will adjust for inflation by using classification for price range during a phone's time frame.

Related Work

First, Nivitus conducted a study in which he used regression analysis to study trends in cell phone pricing during the coronavirus pandemic[2]. By web scraping, he created a dataset that focused on the following characteristics:

- Brand
- Rating
- RAM
- Internal Memory
- Size
- Primary Camera
- Selfie Camera
- Battery Power
- Price

This study is a great introduction into the analysis of cell phone pricing. However, very few predictors were used to cover different aspects of the price. For example, 6 variables were used for hardware characteristics while 2 focused on the social aspect of the device. Also, minimal error tests were conducted. I plan on doing a similar study; however, my study will focus on hardware and will control inflation over time by using a classification method.

Next, Al-Shawwa, Abu-Naser and Mohammed Nasser conducted an analysis on modeling the mobile prices by using Neural Networks.[3] Their shallow model correctly classified mobile phones with 96.31% accuracy. Other metrics such as sensitivity, specificity, recall, etc. were ignored. I will attempt approach my analysis a similar way; however, I will avoid using a similar model. Neural Networks make accurate predictions; however, it's difficult to interpret how a decision is made. In our case, explanation is just as important as model accuracy.

Last, Muhammad Asim and Zafar Khan analyzed the prices of cellphones by using both hardware and branding attributes; also, they focused on comparing results from both naïve bayes and decision tree classifiers. The analysis is in depth; however, the authors forgot to account for inflation. In my study, I will account for it by labeling each phone in accordance to pricing of the year the it was released.

Proposed Work

Data and Tools Overview

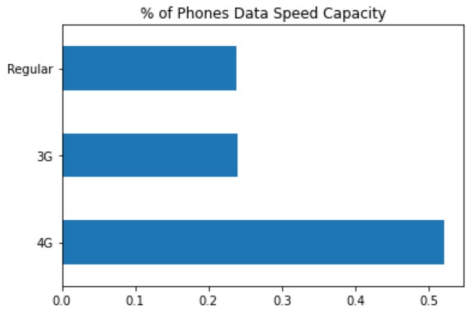
First, I used Python for this entire study. Python is multifaceted and allows us to handle different areas of production. Also, we will use a dataset uploaded to Kaggle by Abishek Sharma[6]. The attributes of our study are:

1. **battery_power** - Total energy a battery can store in one time. Measured in mAh
2. **blue** - Whether the phone has bluetooth or not
3. **clock_speed** - Speed at which microprocessor executes instructions

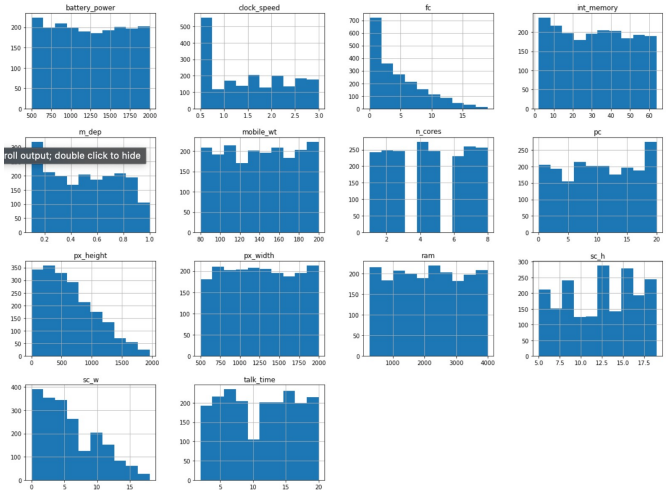
4. **dual_sim** - Has dual sim support or not
5. **fc** - Front camera mega pixels
6. **four_g** - Whether the phone has 4G or not
7. **int_memory** - internal memory in gigabytes
8. **m_dep** - Mobile Depth in cm
9. **mobile_wt** - Weight of mobile phone
10. **n_cores** - Number of cores of a processor
11. **pc** - Primary Camera in mega pixels
12. **px_height** - Pixel Resolution Height
13. **px_width** - Pixel Resoulution Width
14. **ram** - Random Access Memory in Megabytes
15. **sc_h** - Screen Height of mobile in cm
16. **sc_w** - Screen Width of mobile in cm
17. **talk_time** - longest time a single battery charge will last when you are talking
18. **three_g** - Whether the phone has 3G or not
19. **touch_screen** - Whether the phone is touch screen or not
20. **wifi** - Whether the phone has wifi or not
21. **price_range** - Response variable, whether the phone is expensive or not

Data Understanding

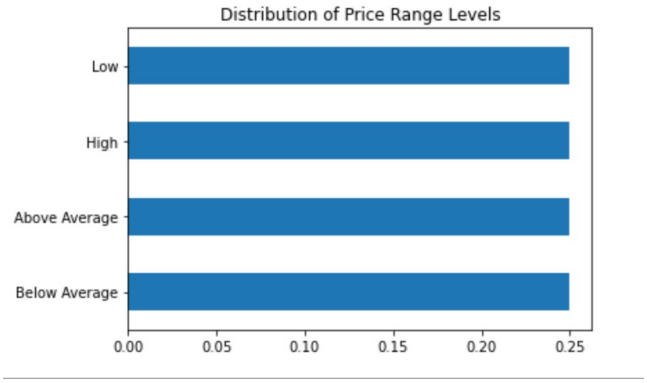
Overall, we have 2000 observations. Of the 21variables, 7 are categorical, including the response variable (price range). All the categorical variables are binary except for the response. Price Range was evenly split between “Low”, “Below Average”, “Above Average” and “High” The levels of most of the categorical variables were evenly split between “Yes” and “No”. However, the distribution of phones’ data speed was unbalanced. About half of the phones have 4G data.



None of our continuous variables follow a normal distribution; In fact, our front camera attribute follows an exponential distribution. Similarly, we will find less phones as the pixel height and screen width increase. Due to the abnormality of the data, I normalized the attributes during the preprocessing stage for the necessary classifiers.



Last, our response variable has an even distribution. When creating our model, this will allow for easier adjustment of any hyper parameters.



Data Preprocessing

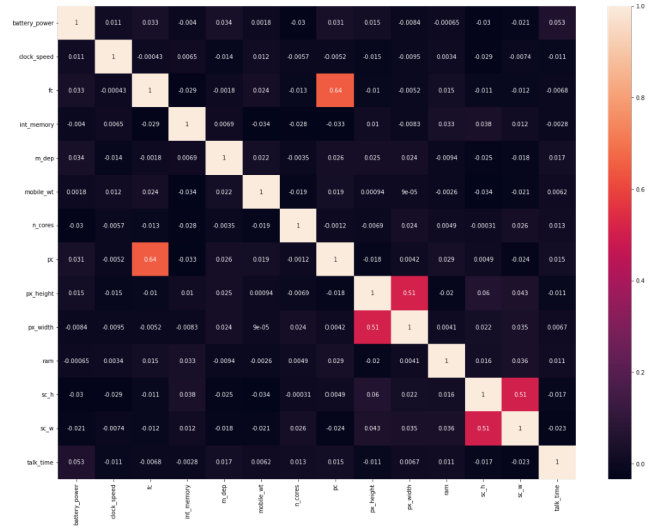
None of our data had missing values; however, some observations, regarding pixel height, were concerning. Some observations contained pixel heights of close to 0 with a large screen width. I decided to remove these data points as they are minimal and inaccurate.

Warehousing

Warehousing was not required for analysis. All our data came from a csv file of 2000 observations.

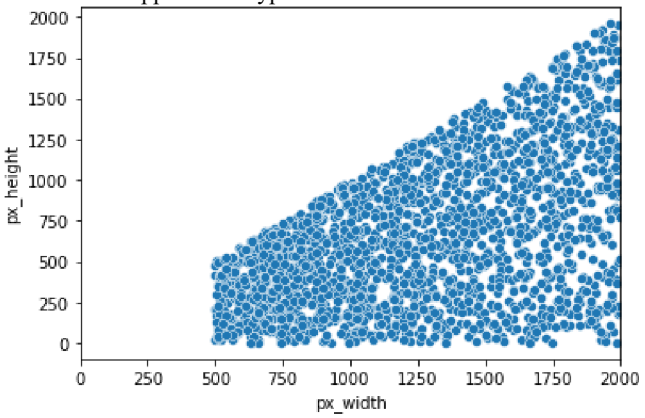
Exploratory Data Analysis

First, I evaluated the relationship of all the continuous variables.

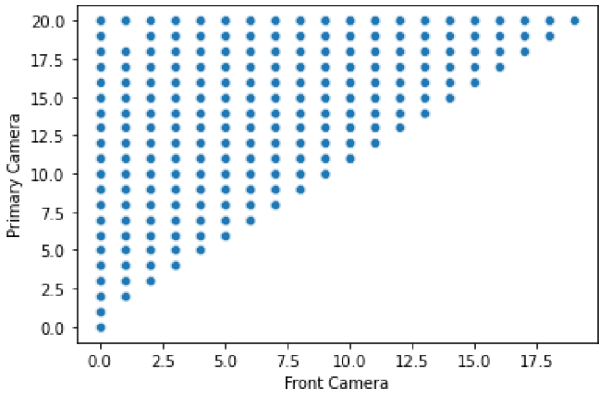


We can see the primary camera and front camera have a high, positive correlation (light red box). This makes sense as many mobile devices have cameras which are similar in megapixels. Also, pixel height and pixel weight are highly correlated. Again, this makes sense as many phones follow a similar resolution makeup.

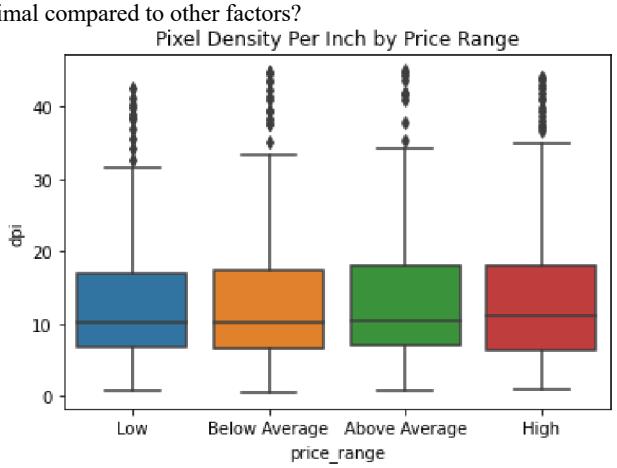
Although, many cell phones have similar pixel resolution, their orientation may differ. For example, some devices are designed to be held vertically while others are not. Our data suggests, for phones with a width larger than 500 pixels, the pixel height will be the same or less. Standard display resolutions are (width x height) 1280 x 1024, 1920 x 1080 and 2560 x 1440. Therefore, our visualization supports our hypothesis.



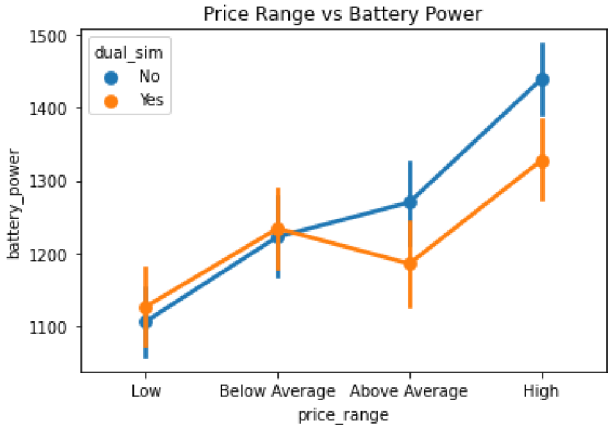
Like our previous analysis, the megapixels in the primary camera are highly correlated to the megapixels in the front camera. The primary camera is either the same or of higher quality than the front camera.



However, many companies measure their resolution by PPI (Pixels Per Inch or Pixel Density). Ideally, a higher PPI means better picture quality. Surprisingly, our data doesn't show a difference of PPI between price ranges. Maybe the difference is minimal compared to other factors?

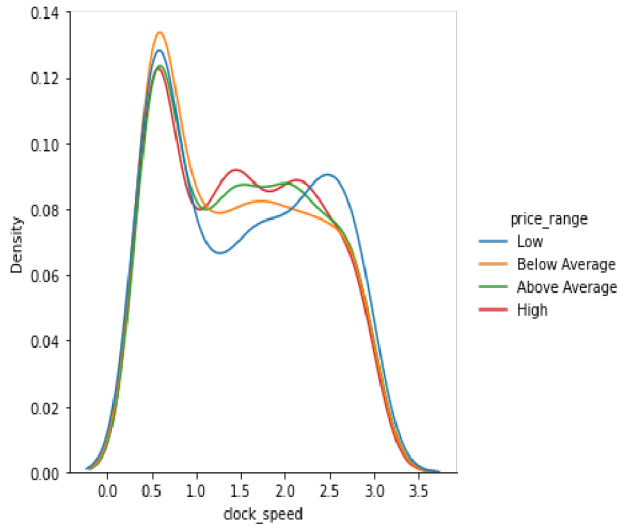


Whether someone is at home, at work or in the woods, many people desire a longer battery life. According to our data, high-priced phones have more battery power than any other price range; on the other hand, low-priced phones have a weaker battery life.

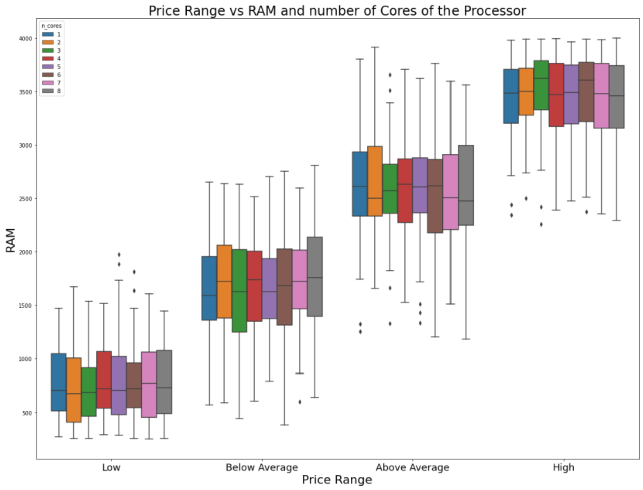


Also, for higher priced phones, dual sim cards are less prevalent. Sim cards are used to store data such as contacts, messages, etc. However, quality smartphones use WIFI, data or internal memory to store information.

Along with more storage, many believe quality phones should perform quickly. However, we find that all but lower priced phones have similar clock speeds. In fact, below average phones seem to perform slightly faster.



Other hardware associated with speed are the number of cores in the processor and the number of gigabytes of random access memory (RAM) in the phone. Typically, one would expect a high-priced phone to have higher RAM and more processing cores. We see the hypothesis is confirmed when discussing RAM; however, the number of cores is not a direct reflection of the amount of RAM in the device.



Touch screen capability of phones creates easy access for everyday tools such as email, social media, and entertainment. To test if touch screen capability (Whether it has it or it doesn't) has any effect on price range, I ran a chi squared test to evaluate the independence of the two factors. The hypothesis is the following:

H_0: Price Range and Touch Screen Capability are independent

H_1: Price Range and Touch Screen Capability are not independent.

touch_screen	No	Yes
price_range		
Above Average	265	235
Below Average	239	261
High	252	248
Low	238	262

Contingency Table for touch screen and price range.

The test resulted in a p-value of 0.275. Therefore, we can't state touch screen capability and price range are dependent.

Evaluation

First, I divided the data into training, validation, and test sets. 70% of the data was used for the training set, 10% for the validation set and 20% for the test set.

When building a classifier for our data, we needed to emphasize interpretability over accuracy; however, we still would like a high-performing model. Thus, I could not use Neural Networks as a potential classifier. The candidate models and their accuracy on the training data were:

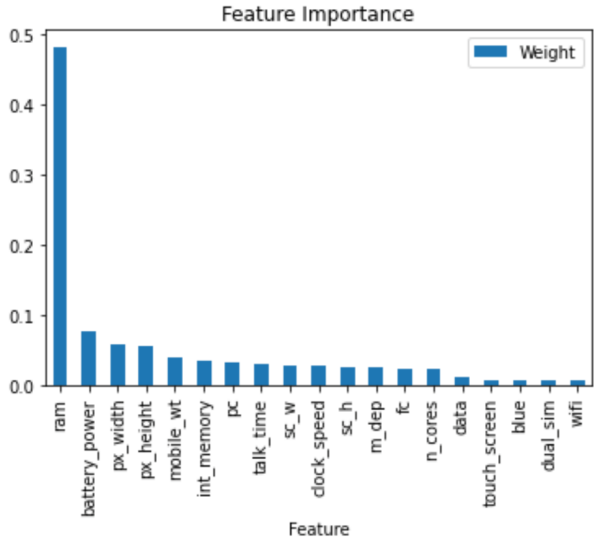
Model	Accuracy (%)
Support Vector Machine	94.56
K Nearest Neighbors	91.56
Random Forest	86.69
Decision Tree	81.69
Multinomial Logistic Regression	71.00

Accuracy is the best evaluation metric for our study. Many customers are apprehensive when purchasing an expensive phone. If our model misclassifies an inexpensive phone, then it may never sell. Also, companies may deter from selling that model. On the other hand, misclassifying an expensive phone can cause a company to lose profits.

From our candidate models, both support vector machine and k nearest neighbors performed well. Although the results can be interpreted, it is difficult to discover the important features using

both models. Due to its simple interpretability and good fit, the best candidate for this study was the Random Forest Classifier.

First, I wanted to review the important features within the model.



The random-access memory (RAM) of a cell phone seems to account for 50% of the decision making within our forest. However, the data type, touch screen capability, wifi, Bluetooth feature nor dual sim capability seems to contribute much information. Therefore, they were dropped from the model. Also, we can see the battery power and pixel dimensions contributed a lot of information to the phone price. This is understandable as many consumers want a phone they don't have to charge often. Also, better pixel density allows for better phone features.

After removing the features, I tested the data on the validation set. Our model had the following results:

	precision	recall	f1-score	support
Low	0.88	0.93	0.91	88
Below Avg	0.88	0.86	0.87	108
Above Avg	0.83	0.88	0.85	97
High	0.95	0.88	0.91	107
accuracy			0.89	400
macro avg	0.89	0.89	0.89	400
weighted avg	0.89	0.89	0.89	400

Regardless of having imbalanced classes within our validation set, our model performed well for all classes. the accuracy of our classifier improved 3%. However, our model has varying precision scores. In other words, if our model identifies a phone as 'high priced', then it is more likely true compared to 'above average priced' devices. Regardless, our overall f1 score, which evaluates the harmonic mean of our model's precision and recall, is very high. Also, low and high-priced phones are more accurately classified than below average and above average ones. This makes sense as both price ranges are on the ends of the spectrum; therefore, the features that put them in the price range are more distinct than the two average ranges.

Next, I used the 'GridSearchCV' method within the sklearn module to tune hyperparameters. Adjusting these parameters helped me improve the model fit. More information about the hyperparameters of Python's Random Forest Classifier can be found in there Sklearn module API resource. The following were adjusted (final decision shown in parentheses):

- bootstrap (True)
- criterion (entropy)
- max_depth (10)
- min_samples_split (5)
- n_estimators (500)

These adjustments improved the precision of above-average priced phones and the recall of high-priced devices.

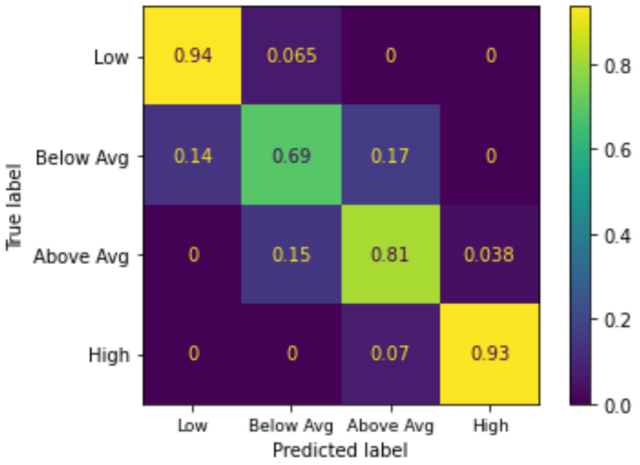
	precision	recall	f1-score	support
Low	0.88	0.94	0.91	88
Below Avg	0.88	0.86	0.87	108
Above Avg	0.86	0.86	0.86	97
High	0.94	0.92	0.93	107
accuracy			0.89	400
macro avg	0.89	0.89	0.89	400
weighted avg	0.89	0.89	0.89	400

Final Model Evaluation

After adjusting our model, I noticed our accuracy didn't improve. Therefore, I finalized the model and evaluated its performance on the test set.

	precision	recall	f1-score	support
Low	0.91	0.94	0.92	62
Below Avg	0.71	0.69	0.70	42
Above Avg	0.81	0.81	0.81	53
High	0.95	0.93	0.94	43
accuracy			0.85	200
macro avg	0.84	0.84	0.84	200
weighted avg	0.85	0.85	0.85	200

Compared to our validation set, our model's accuracy slightly declined. This may be due to a small amount of overfitting. Our classifier is very accurate for all classes except below-average priced phones.



As shown by the confusion matrix above, 69% of the below-average priced phones were labelled correctly; however, 1 of every 5 were labelled as above-average priced. On the other hand, above-average priced phones were labelled as below-average at a similar rate. Thus, below-average phones may have more similarities with above-average phones than other price ranges.

Discussion

First, we have completed this project within the given time frame. All the following checkpoints have been completed:

- January 17: Have Proposal Completed
- January 24: Have all data collected
- February 2: Finish Cleaning and Combining Datasets
- February 15: Finish all Preprocessing and Outstanding Visualizations
- February 19: Finish Comparing Potential Models
- February 22: Tune Final Model and Test
- February 27: Submit project with presentation

Overall, the project was completed with minimal issues. Initially, I wanted to scrape for data regarding brand names and location; however, I struggled to find valid data within the short time frame. Given the 8-week time frame, I struggled to complete a deeper analysis. However, this project may be extended to include more features.

Conclusion

Cellphones have evolved to become vital in our everyday lives. For both companies and consumers to make smart choices, they must understand how cell phone hardware affects pricing.

Using their characteristics, I identified that cell phones have indistinguishable differences. Expensive phones tend to have more RAM with longer battery lives. Also, a phone's datatype, touch-screen capability, WIFI feature, Bluetooth availability and dual-sim card have a minimal effect on their pricing. Last, below-average and above-average priced phones are more difficult to distinguish from each other than other price ranges.

For future work, others can extend this project by incorporating other factors such as a phones brand. Some brands such as Samsung and Apple tend to sell expensive phones. On the other hand, companies such as Motorola and LG tend to manufacture low-cost devices. Controlling for brand name may help identify how different companies produce their phones. Also, one can conduct a temporal study on cell phone trends. Recently, phones have become much faster. What has caused this trend. How has the hardware and/or software improved? Knowing the value of your phone is important. So, next time you reach for your phone, think to yourself. Was your phone worth the price you paid for it?

References

[1] G., Deyan. (2022, February 6). *How Much Time Does The Average American Spend on Their Phone in 2022?*. Techjury. https://techjury.net/blog/how-much-time-does-the-averageamerican-spend-on-their-phone/#gref

[2] Nivitus. (2020, July 22). *Mobile Price Prediction Using Machine Learning.* Medium. https://medium.com/@Nivitus./mobile-price-predictionusing-machine-learning-fa9cab6fb242

[3] Al-Shawwa, Mohammed; Abu-Naser, Samy S. and Mohammed Nasser, Ibrahim. (February 2019). *Developing Artificial Neural Network for Predicting Mobile Phone Price Range.* ResearchGate. https://www.researchgate.net/publication/331398317_Developing_Artificial_Neural_Network_for_Predicting_Mobile_Phone_Price_Range

[4] Asim, Muhammad and Khan, Zafar. (2018, March). *Mobile Price Class Prediction Using Machine Learning Techniques.* Towards Data Science. https://www.ypulse.com/article/2019/10/01/how-gen-zmillennials-are-discovering-music-on-social-media-in-2charts/

[5] Zeman, Eric. (2020, July). *Phone Prices Have Almost Doubled Over the Past Five Years. Why is That?.* Android Authority. https://www.androidauthority.com/how-smartphone-priceshave-changed-1134574/

[6] Sharma, Abhishek. *Mobile Price Classification*. Kaggle. https://www.kaggle.com/iabhishekofficial/mobile-priceclassification