Mobile Phone Price Analysis and Prediction

https://github.com/AllHailTheSheep/CIS3715/tree/main/final

Julia Fasick
CIS 3715: Principles of Data Science
Hongchang Gao
April 28, 2025

Introduction

In recent years, mobile devices have both surged in cost, and become an absolute necessity. While phones were once considered a luxury, the onset and normalization of texting, social media, and "always online" culture have made them essential for most. At the same time, the price of these devices has skyrocketed since the early years of smartphones. In 2008, the Apple iPhone 3G launched for $199 MSRP (*Apple Introduces the New IPhone 3G*, 2017), which is an unimaginable price in today's mobile device economy, particularly for a brand new flagship device. This raises the relatively new challenge of buying a device that is deemed absolutely necessary for your personal success, and being fiscally responsible. According to a 2023 report on the economic wellbeing of American households, 13% percent of Americans reported they would be unable to cover a $400 emergency cost, whether by putting it on a credit card, borrowing from a friend, taking out a loan, selling something, or taking a payday loan (*Expenses*, 2023). Ideally, such a purchase should be budgeted for long in advance, but this raises a new question: how much should one put aside for a mobile device? With the increase of cost in recent years, it seems unlikely the same amount that bought a flagship phone this year will do the same in a few years time. This paper attempts to answer that question by using a linear regression with ℓ^2-normalization in addition to a random forest regressor with one hundred decision trees to predict the price of upcoming devices based on it's specifications, such as manufacturer, RAM, battery capacity, and screen size. We will then apply these findings to find the amount per month that should be put aside to comfortably afford a flagship device when it comes time to upgrade. This paper includes both quantitative data analysis in combination with qualitative insights gained from working in the mobile phone repair business for over four years.

Approach

A linear regression will be utilized to predict the price of the devices. Features such as higher RAM are associated with a higher cost. A regression can learn these features and extrapolate upon it, allowing us to predict the cost of future devices. Several challenges could throw the linear regression off. Specifications such as model name and processor name mean nothing to the regression, even once encoded with an ordinal encoder, as no models share model names and very few share processors. These features contribute nothing to the regression, in fact it acts as a detriment to the linear regression. Additionally, certain features, like higher amounts of RAM, will be more common at a cheaper price after a few years. With a regression trained on data showing up to 16GB devices, a 16GB device down the line could cost a significant amount less than what the regression predicts. Ideally, ℓ^2-normalization will help overcome this feature shift by reducing dependency collinear features, allowing the regression to "learn around" the fact that many features steadily increase at a relatively similar rate, but the overall effect on price is largely dependent on the release year.

In addition, a random forest regressor will be used. The random forest regressor will consist of one hundred individual decision trees to allow the regressor to learn some of the variance in the dataset. For example, some of the devices are tablets, therefore very heavy, which influences price in ways more complex than what a linear regression can handle. Additionally, certain companies such as Apple or Samsung have a much more premium price tag attached to them. The linear regression cannot learn what the company differences indicate as well as the random forest regressor can.

The dataset needed significant preprocessing before use. Most features were strings, and needed some regular expressions to convert into a numerical value. Some features, like the front camera column, had specific values that needed to be preprocessed carefully. Some devices are listed as "12MP", others as "12MP / 4K", and others still as "Dual 12MP". The column had to be iterated through and scanned for each of these patterns before extracting the numerical features. The back camera feature had similar values, in addition to values such as "12MP + 12MP + 12MP" for many Apple devices. For simplicity's sake, only the first value was used, as it was usually the most representative. The linear regression may get thrown off by this, but since many company's devices have similar values for these features, the

random forest regressor should not have an issue learning that a value of “12” for the back camera on an Apple device is more indicative of a higher cost than for a Oppo or Vivo device. The RAM feature also often had multiple values, such as “8GB / 12GB”. This means that with a higher storage size, the device comes with more RAM. Since the price listed is always for the base model, the first value is used. There was also a company with repeat names (“POCO” and “Poco”) which had to be merged before label encoding.

The dataset contains 930 samples of devices across 18 companies released over 11 years. Figure 2.1 and 2.2 show the distribution of devices across the companies and launch years. Note the most represented company is Oppo, a Chinese technology company not often seen in the United States of America. They are known for releasing many models targeted at many different niches, so it makes sense that they are the most represented (*Oppo Smartphones*, 2024). Figure 2.3 shows the features for use in the regression after preprocessing of the raw data, and Figure 2.4 shows the principle component analysis of the dataset in two dimensions.

There are several interesting conclusions that can be drawn from the dataset before the regression is even trained. Figure 2.5 shows the median price per year of the devices in the dataset. The rapid price increase from 2014 to 2018 is indicative of how fast smartphones caught on. Interestingly, after that, there is a distinct downward trend in price. This most likely represents the introduction of budget devices. Parts, licensing, manufacturing, and 5G connectivity are all cheaper than ever, resulting in budget devices with considerable power and features for a lower price point (Stanton et al., 2022). Another interesting conclusion can be reached when plotting the median RAM per year for the samples, as in figure 2.6. The average amount of RAM skyrockets from 1.5GB to 4GB from 2014 to 2016. The rise of RAM initially comes close to following Moore’s Law, doubling just about every two years (*Moore’s Law*, 2023). Eventually, the RAM plateaus at 8GB. This is generally a good value for a smartphone, as it’s enough to allow multitasking without any slowing down. Additionally, in recent years, software gains and other optimizations have provided gains to performance as well (Hildenbrand, 2020), although in coming years the onset of artificial intelligence applications in mobile devices may require an increase in resources.

Results

The ℓ2-normalized linear regression was able to achieve a MAE of 246.2069, a MSE of 108405.6680, and a RMSE of 329.2502. These values indicate the regression will be off by an average of ~250$ for any given device. Most likely, the issues talked about above with collinear features and preprocessing removing some feature values are throwing the regression off.

The random forest regressor was able to provide a much more accurate prediction. It achieved a MAE of 84.9795, a MSE: 14746.5354, and a RMSE: 121.4353. This means that the prediction will be off by ~85$ on average. While this seems like a lot, in terms of many of the flagship devices, that is a relatively small percentage of the cost. Additionally, the regressor performs a lot more accurately on budget devices, meaning that the average error usually only appears on those flagship devices, as well as devices that are outliers in their companies.

Figure 3.1 shows the performance of both the regressions on seven devices. The first two (Samsung S25 Ultra and Apple iPhone 16e) are not in the dataset, and the other devices are present in the dataset. The linear regression clearly did not fit very well, even estimating a negative price for the budget oriented Vivo Y12s. The random forest regressor performs quite well, with the notable exception being the Apple iPhone 16e, which is a large outlier in terms of its price for Apple devices. Figure 3.2 shows the predictions for several upcoming devices. These devices are not out, so only time will tell how accurate the prediction is, but metrics and intuition from previous behavior from the random forest regressor tells that most random forest predictions should be relatively accurate (+/-$100) and may be estimating low on the upcoming S25 Edge. This would be due to the fact the Edge is a reintroduction of an older series of devices, meaning the regressor does not have many modern similar devices from Samsung to draw the decision tree from. The decision trees can be plotted, and part of the first decision tree the regressor uses

is included in Figure 3.3. The full plotting of this tree can be found on GitHub, but is too large to include here.

Conclusion

From our prediction of upcoming devices using both linear regression and a random forest regressor, it is clear that the price range of future devices should remain in the $1,000 - $1,400 for an ultra high end device. Alternatively, a new base model flagship device will continue to be priced around $800. This, of course, is barring some unexpected variable changing, such as a trade war with a county that produces most of these devices. This is not an increase from the current price trends, which demonstrates a plateauing of mobile device prices. This can even be seen starting over the last few years, with devices like the Samsung Ultra series launching consistently around $1,200 for the last few models, and the base model of the Apple iPhones launching at $799 since the iPhone 12. Samsung has followed suit, using the $799 base model price tag from the S22. This is great news for consumers, as the challenge of budgeting for a device like this is slightly abated. The question still remains of how much should be put aside per month to comfortably afford such a device. The average device is replaced about every 2.5 years (Laricchia, 2023), but I suggest that budgeting for a single battery replacement will allow devices to get four years of life without an issue. As the person who does those battery replacements, the price of a battery replacement is usually around $100. Assuming you do not need the most powerful device on the market, this would mean that you will want $900 to afford a base model device, or $1,300 for the ultra high end flagship. This leads us to the conclusion that over 4 years, if you are just looking for a base model device, you only need to put away $18.75 a month. The people who care for the higher powered devices should be putting away $27.08 a month. With these numbers, affording a new device when the time comes should not be an issue, all for the price of a coffee a week.

Appendix

Figure 2.1 shows the distribution of samples across multiple companies.

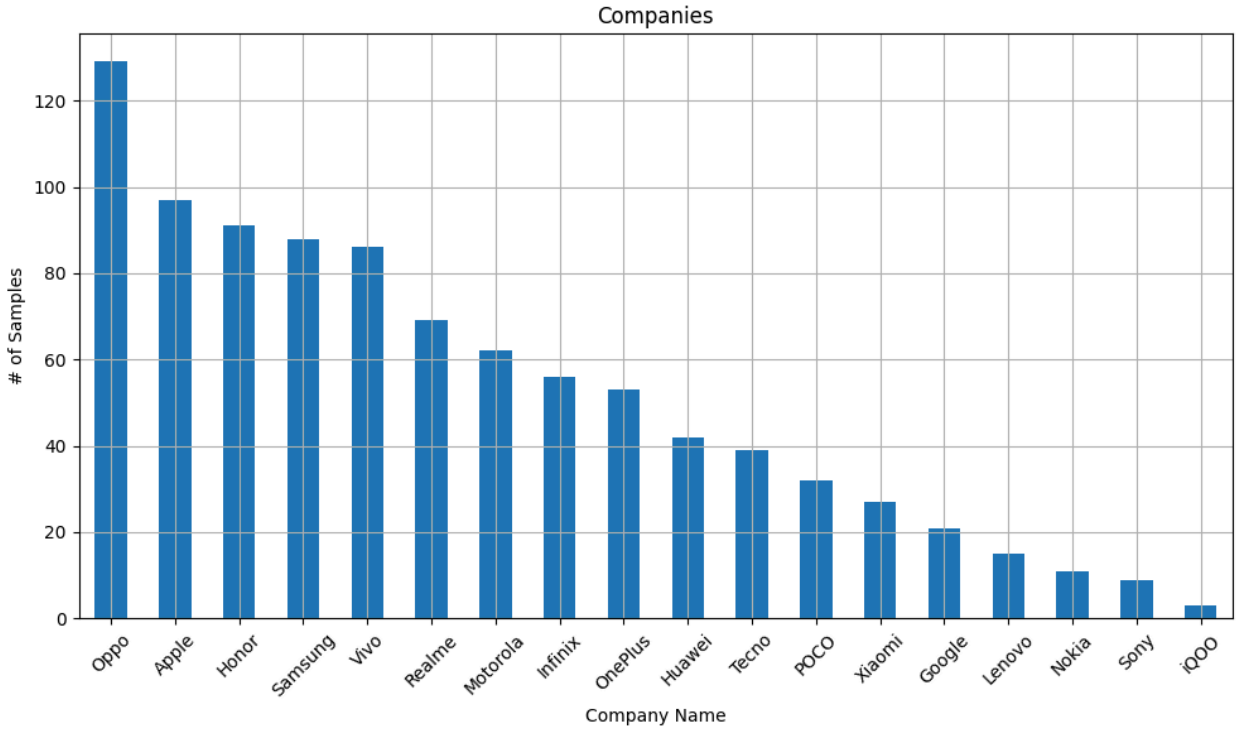


Figure 2.2 shows the distribution of samples in the dataset across their launch years.

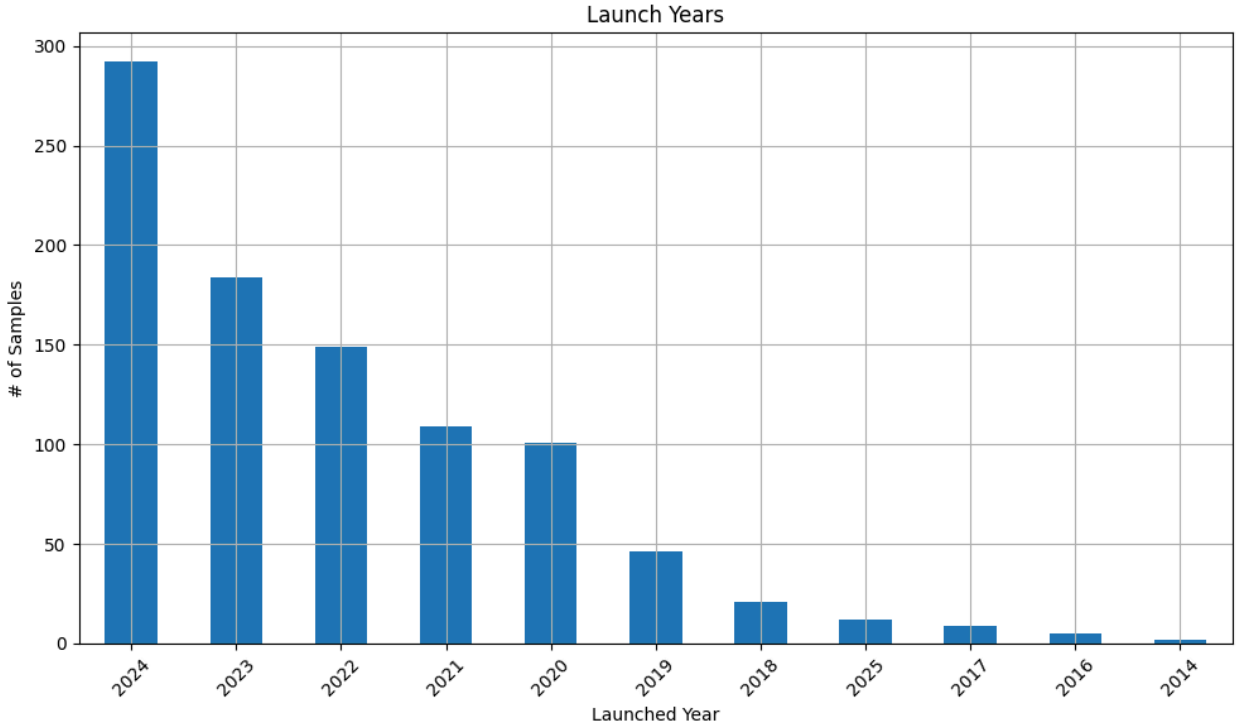


Figure 2.3 shows the features in the data for use in the regressions after preprocessing, as well as the target launch price.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 930 entries, 0 to 929
Data columns (total 9 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Company Name          930 non-null    int64
 1   Mobile Weight         930 non-null    float64
 2   RAM                   930 non-null    float64
 3   Front Camera          930 non-null    float64
 4   Back Camera           930 non-null    float64
 5   Battery Capacity      930 non-null    int64
 6   Screen Size           930 non-null    float64
 7   Launched Price (USA)  930 non-null    float64
 8   Launched Year         930 non-null    int64
dtypes: float64(6), int64(3)
memory usage: 65.5 KB
```

Figure 2.4 shows the principle component analysis of the dataset into two dimensions.

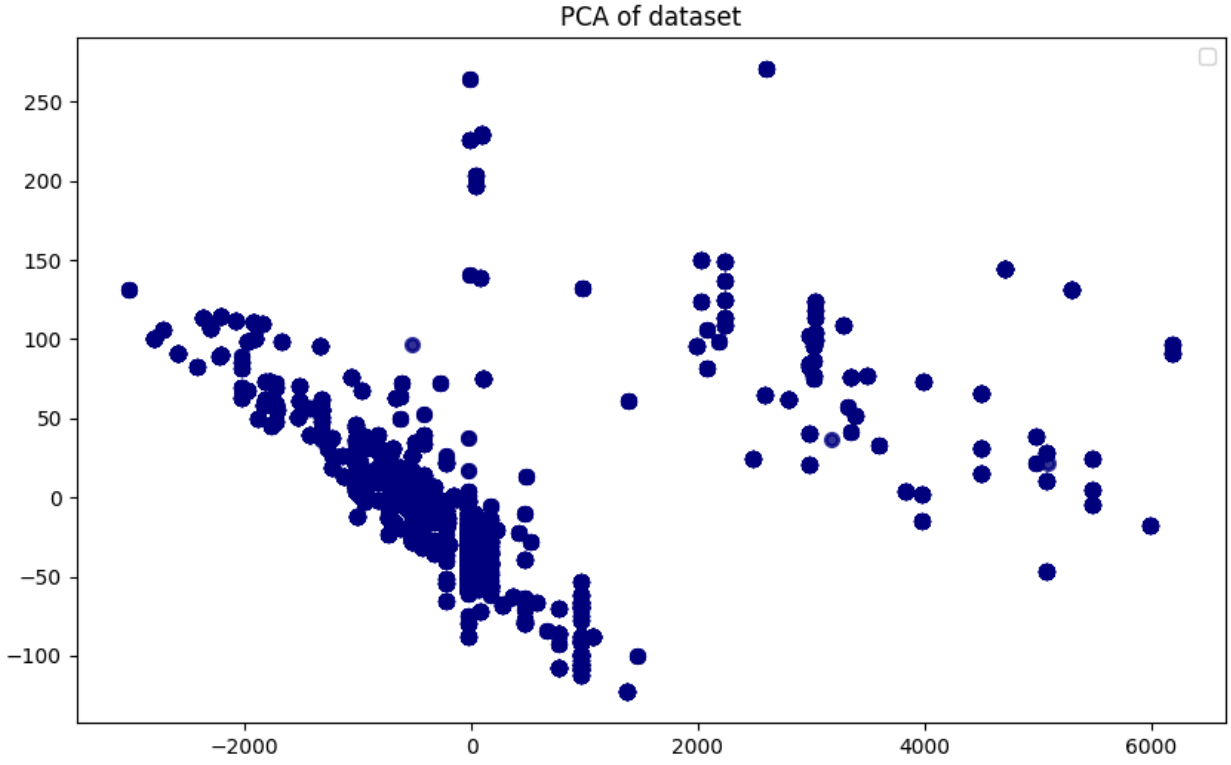


Figure 2.5 shows the median price of the samples by launch year.

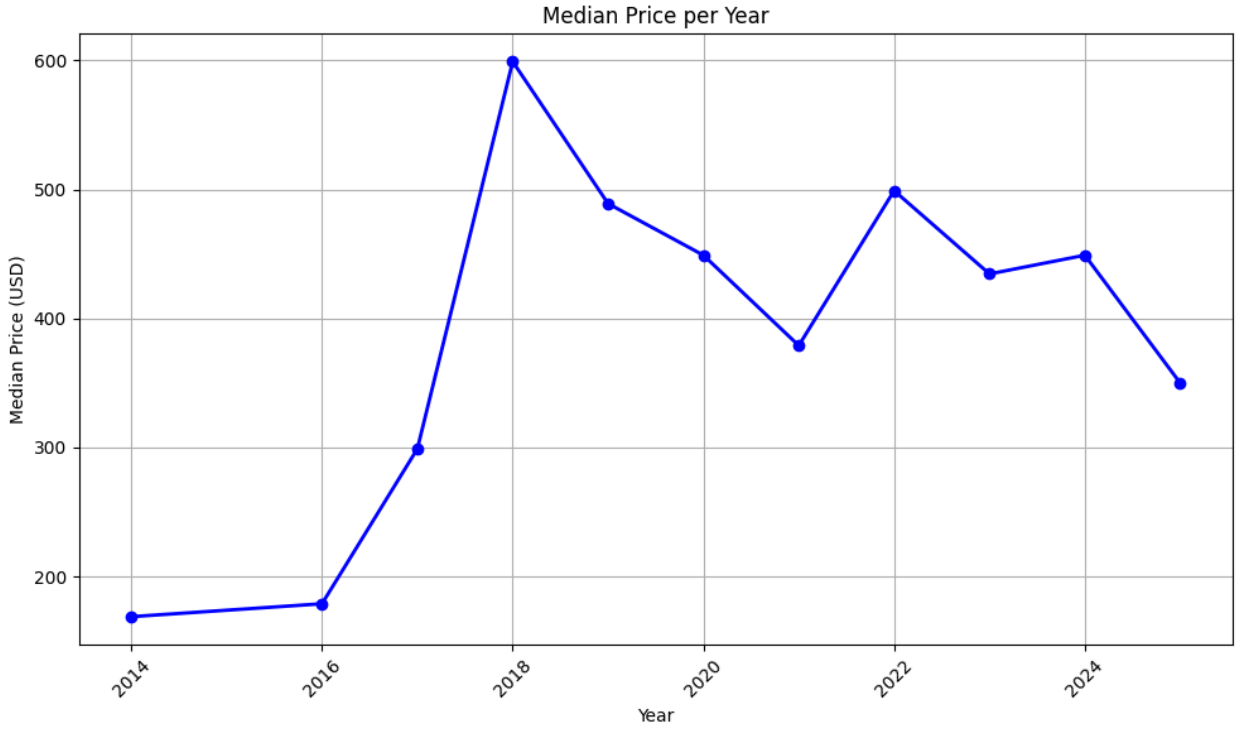


Figure 2.6 shows the median amount of RAM in the dataset per launch year and the initial correlation with Moore's Law.

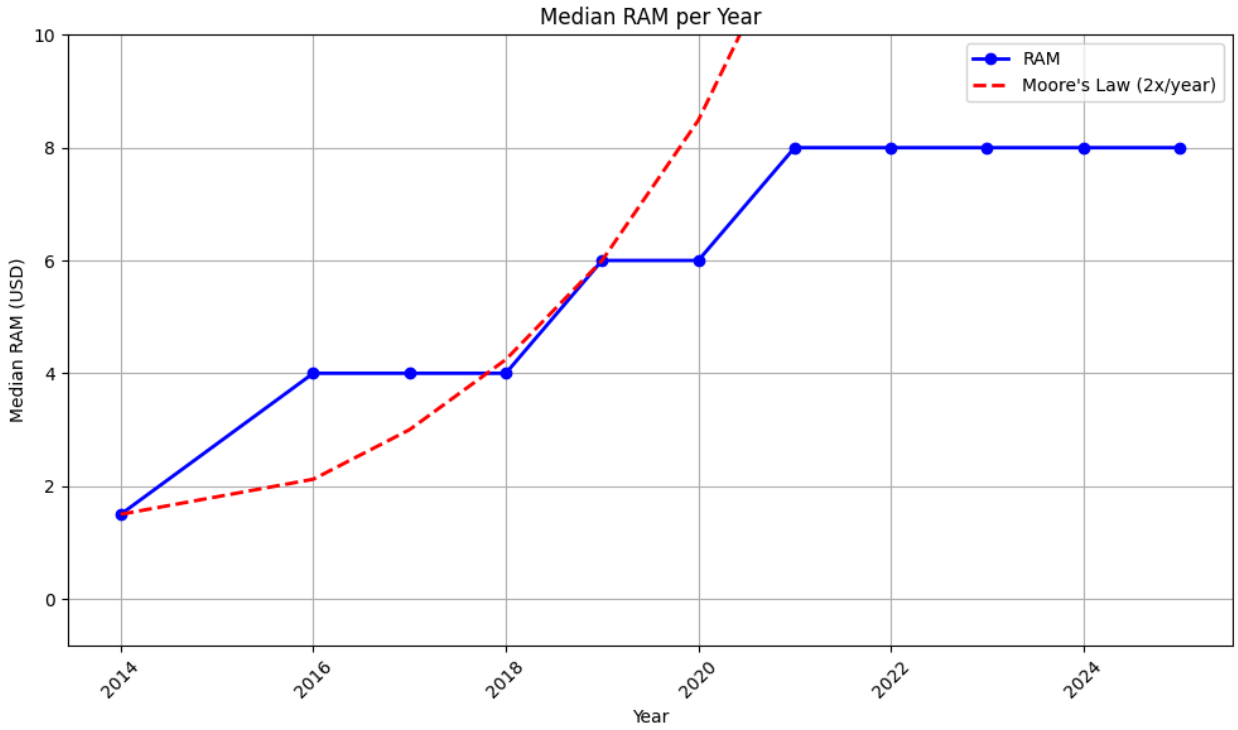


Figure 3.1 shows the performance of both the linear regression and the random forest regressor on several devices. Two of the devices, (the Samsung S25 Ultra and the Apple iPhone 16e) are not in the dataset, and the other devices are present in the dataset.

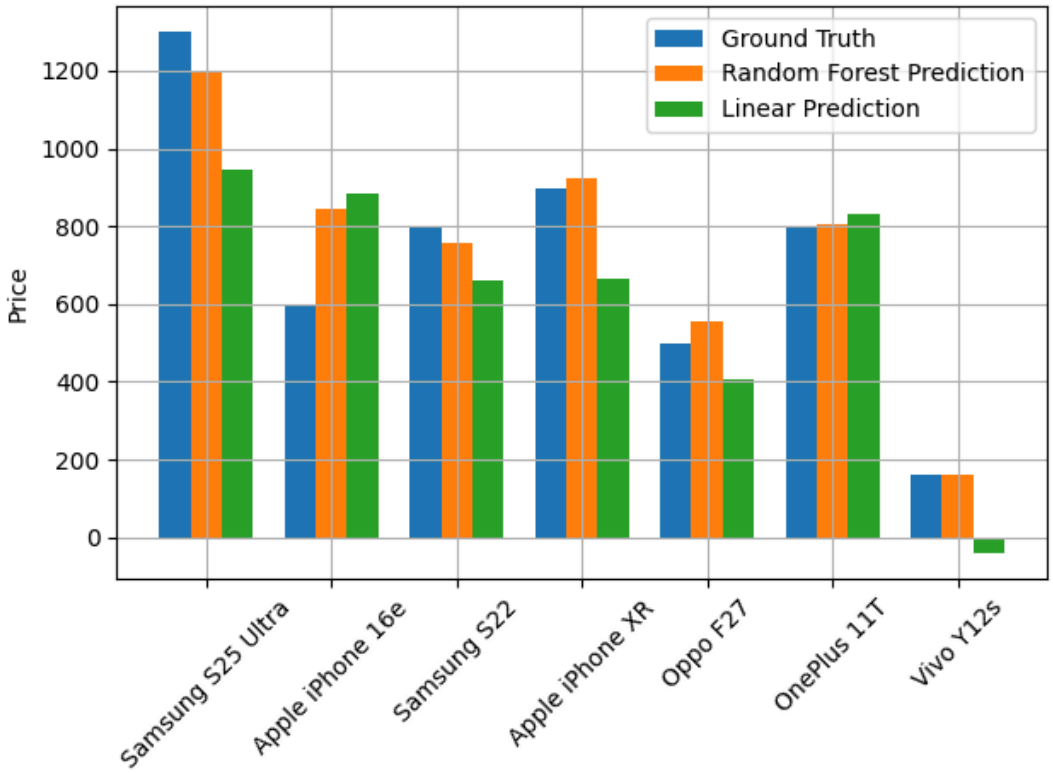


Figure 3.2 shows the performance of both linear regression and the random forest regressor on several upcoming devices.

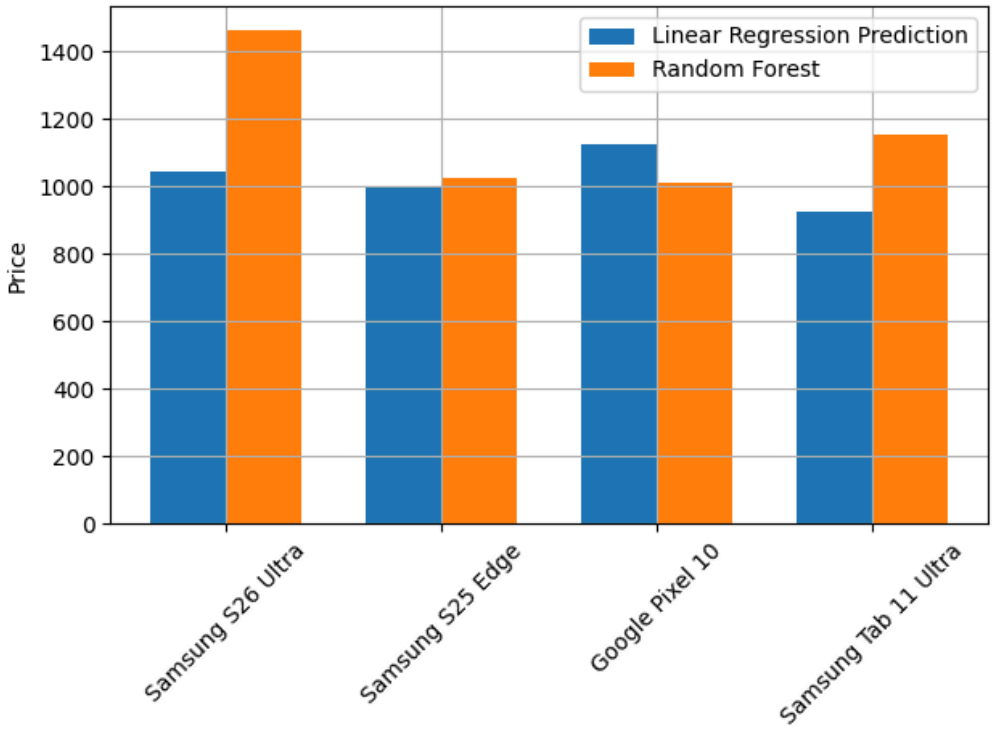
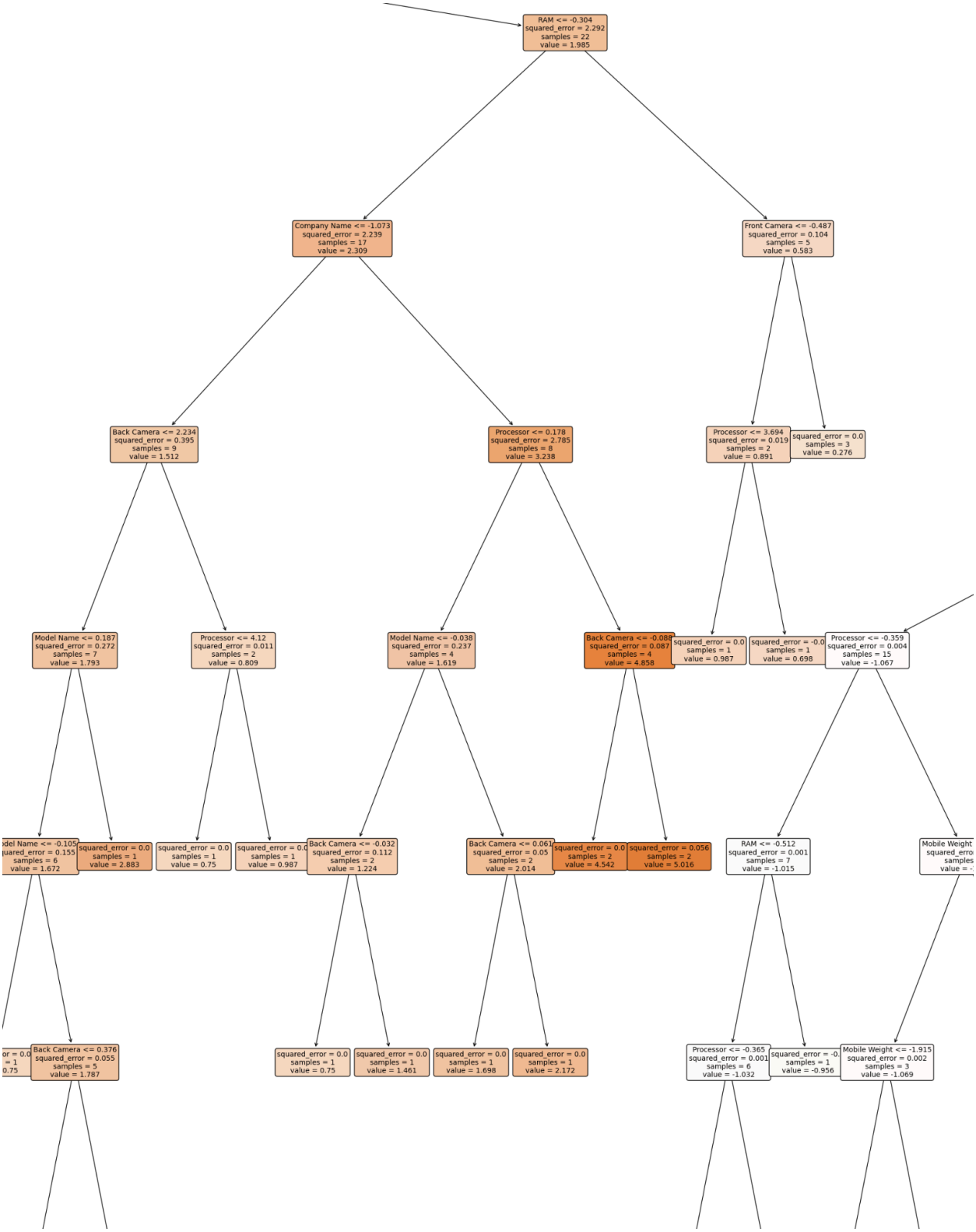


Figure 3.3 shows part of a decision tree used by the random forest regressor.



Works Cited

Dataset: https://www.kaggle.com/datasets/abdulmalik1518/mobiles-dataset-2025/data

Samsung Galaxy Tab S11 Ultra All Specs and Price. (2025). Specs-Arena.com. https://specs-arena.com/samsung-galaxy-tab-s11-ultra/

Apple Introduces the New iPhone 3G. (2017). Apple Newsroom; Apple. https://www.apple.com/newsroom/2008/06/09Apple-Introduces-the-New-iPhone-3G/

Diaconescu, A. (2025, April 25). *Convincing new leak essentially confirms many of Samsung's key Galaxy S25 Edge selling points*. PhoneArena. https://www.phonearena.com/news/samsung-galaxy-s25-edge-key-specs-camera-weight-thinness-materials-confirmed_id169808

Expenses. (2023, May). Www.federalreserve.gov. https://www.federalreserve.gov/publications/2023-economic-well-being-of-us-households-in-2022-expenses.htm

Google Pixel 10 specs (Rumored) - PhoneArena. (2025, April 26). PhoneArena. https://www.phonearena.com/phones/Google-Pixel-10_id12651

Hildenbrand, J. (2020, April 19). *RAM: What it is and when do you need more?* Android Central. https://www.androidcentral.com/ram-what-it-how-its-used-and-why-you-shouldnt-care

Laricchia, F. (2023, February 28). *Smartphones replacement cycle in the US 2014-2024*. Statista. https://www.statista.com/statistics/619788/average-smartphone-life/

Moore's Law. (2023, September 6). GeeksforGeeks. https://www.geeksforgeeks.org/moores-law/

OPPO Smartphones, Top Smartphones | OPPO Global. (2025). OPPO Smartphones; OPPO. https://www.oppo.com/en/smartphones/

Preslav Kateliev. (2025, February 26). *Samsung Galaxy S26 Ultra release date expectations, price estimates, upgrades*. PhoneArena. https://www.phonearena.com/galaxy-s26-ultra-release-date-price-features-news

Stanton, B., Lee, P., Wigginton, C., & Hofmeyr, G. (2022, November 30). *Accessible is possible: Introducing the US$99 5G smartphone*. Deloitte Insights. https://www2.deloitte.com/us/en/insights/industry/technology/technology-media-and-telecom-predictions/2023/99-dollar-smartphone-price-trends.html