Project 1: Report/Milestone 3

Jahedur Rahman

Data Science, Bellevue University

DSC 680: Applied Data Science

Professor Catie Williams

September 25, 2022

Mobile Price Prediction

In this project, I predicted the price range of a mobile phone by using the specifications and features of the phone.

Background/History

The price of a mobile phone can differ significantly based on what features the phone has. Presently, the cost can go higher than $1,000. These phones are usually high-end phones with the best specifications. There are also low-end and mid-range phones that don't have the best features. However, they still get the job done.

Business Problem

Companies must be able to predict the price range of a phone they want to make and sell. It can also help them decide their target audience, like if they should sell the device as a budget phone or a high-end phone. This model will be able to predict the price range of a mobile device based on the features and specifications included.

Data Explanation

Dataset

The Mobile Price Classification dataset will be the primary dataset used for this project. This dataset has all the needed columns listed below.

- battery_power - Total energy a battery can store in one time measured in mAh
- blue - Has bluetooth or not
- clock_speed - speed at which microprocessor executes instructions
- dual_sim - Has dual sim support or not
- fc - Front Camera mega pixels
- four_g - Has 4G or not

- int_memory - Internal Memory in Gigabytes
- m_dep - Mobile Depth in cm
- mobile_wt - Weight of mobile phone
- n_cores - Number of cores of processor
- pc - Primary Camera mega pixels
- px_height - Pixel Resolution Height
- px_width - Pixel Resolution Width
- ram - Random Access Memory in Mega Bytes
- sc_h - Screen Height of mobile in cm
- sc_w - Screen Width of mobile in cm
- talk_time - longest time that a single battery charge will last when you are talking
- three_g - Has 3G or not
- touch_screen - Has touch screen or not
- wifi - Has wifi or not
- price_range - 0(low cost), 1(medium cost), 2(high cost) and 3(very high cost)

Data Preparation

There was not a lot of preparation needed for this dataset. Since all the columns were required, no columns were removed. Also, I found no null values so I didn't need to remove any records. I split the dataset into training and testing datasets to fit the models.

Methods

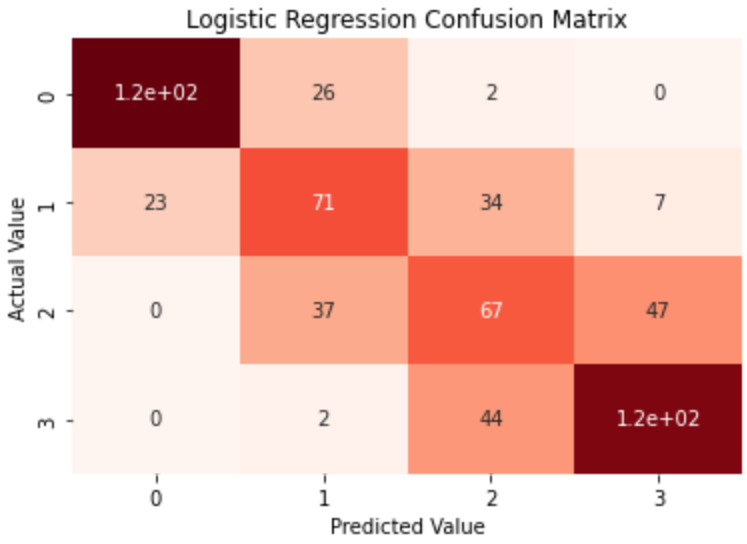
The target variable in the dataset is price_range. There are four price ranges. The price range 0 is a low cost, 1 is a medium cost, 2 is a high cost, and 3 is a very high cost. Since this is a classification scenario, a logistic regression model was used. In addition, other models such as a

random forest classifier, and a decision tree classifier were used to compare the results of the logistic regression.

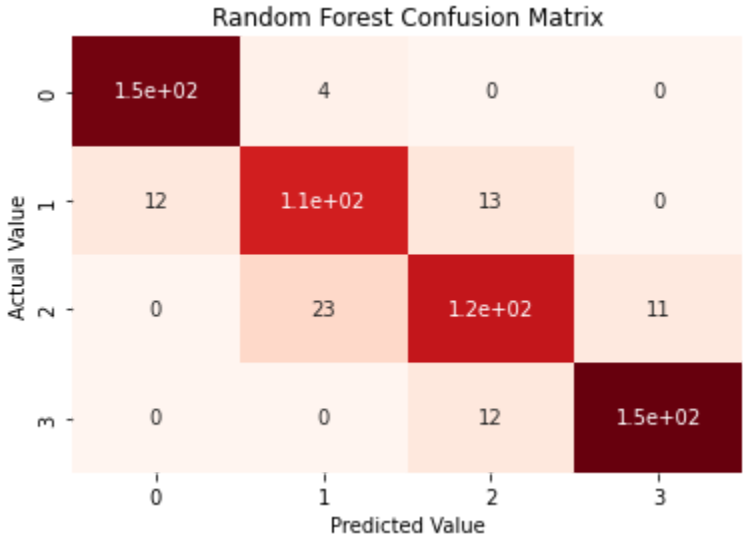
Analysis

To start off with my analysis, I started with a heatmap. I wanted to see if there was any correlation between the variables.

In this heatmap, we can see there is a very high correlation between price_range and ram. This tells us that the amount of ram of a phone has the biggest effect on its price. Next, I built a logistic regression model and computed the accuracy score, and created a confusion matrix.

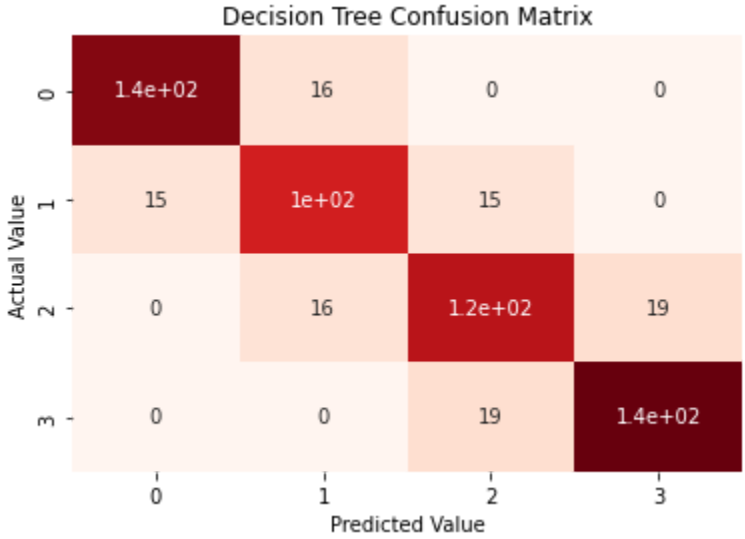


The accuracy score for the logistic regression model was 63%. The confusion matrix shows us that the model correctly predicted 378 out of 600 values and incorrectly predicted 222 out of 600 values. These results didn't seem that good so I built a random forest model and compared the results.



The accuracy score for the random forest model was 87.5%. The confusion matrix shows us that the model correctly predicted 530 out of 605 values and incorrectly predicted 75 out of 605

values. These are much better results compared to the logistic regression model. Next, I built a decision tree model and compared the results.



The accuracy score for the decision tree model was about 83%. The confusion matrix shows us that the model correctly predicted 500 out of 600 values and incorrectly predicted 100 out of 600 values. These results are very close to the random forest model results. However, the results from the random forest model were still better.

Conclusion

Based on the model evaluation results, the random forest model will be used to predict phone prices. It has the highest accuracy and a higher number of correctly predicted values.

Assumptions and Limitations

There is a chance that the dataset has some incorrect information. Since for many of the columns there are values like 0 and 1, wrong values could be accidentally inputted. In addition, the values for these specifications aren't values you would see on high-end phones presently. For example, nowadays there are phones that go way above 4k megabytes in ram.

Challenges

After analyzing the dataset with the heatmap, I feel some features aren't accurately represented. Features such as clock speed, front camera megapixels, and primary camera megapixels should have a bigger effect on the price of a phone. In the real world, many people research these specifications before buying a phone.

Future Uses/Additional Applications and Recommendations

The model can be used to figure out a competitive price to place on a phone. In addition, the model could be used vice versa. A price range can be looked at to figure out what specifications and features a phone should be built with. This could be helpful when targeting specific customers. For example, customers that are on a budget and are looking for a budget phone.

Implementation Plan

The implementation plan is to use this model as an initial decision on the price range of a phone. This will help determine the features and specifications of the final product.

Ethical Assessment

Since the price of a mobile device has an impact on whether a person would buy the device or not, it must be ensured that the price accurately reflects the features of a given device. The mobile device market is very competitive with some phones being sold with a price tag of between $1,200 and $1,500. To compete with these prices, the best price range must be predicted correctly.

References

Sharma, A. (2017). *Mobile Price Classification*. Kaggle. Retrieved September 4, 2022, from https://www.kaggle.com/datasets/iabhishekofficial/mobile-price-classification

Appendix A

Under the Methods section on page 3, it is written, “The price range 0 is a low cost, 1 is a medium cost, 2 is a high cost, and 3 is a very high cost.” A dollar price range was not specified on the site this dataset was taken from. So additional data must be found and research must be done to include an actual price range before this could be implemented in the real world.

Under Ethical Assessments on page 7, it is written, “The mobile device market is very competitive with some phones being sold with a price tag of between $1,200 and $1,500.” The price range specified here is for the prices of phones that are presently found in stores. It does not reflect the price range of any of the phones in the dataset.