Siyu Wang
DATA 606 Spring 2020
Capstone Project Final Report

Mobile Price Range Prediction

Project Overview and Purpose

The main goal of this project is to build a model to predict the price range indicating how high the price is of mobile phone based on battery power, 4G, wifi, Bluetooth, Ram, Internal Memory and other mobile specifications. During the whole project, I will use python to training data, visualize the data, use machine learning algorithm and build models to finish my project. This project will help mobile phone manufacturers predict the reasonable market price of their new phone, thereby strengthening their own market competitiveness. Also, it can help consumers verify dose they paid the best price for their phone. Dose they pay more for the phone itself or for the brand.

Motivation

Once I saw a phone called "Vertu Signature Touch" sell for $15,000. I really want to know what the actual value of this phone is.

Data

The dataset I will use for this project is *Mobile Price Classification.* It is open resource at Kaggle. The dataset contains two csv files. One is train. csv and another one is test.csv. Both files contain 21 columns and 2000 entries. The 21 columns are id(ID), 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 Megabytes), 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), three_g(Has 3G or not), touch_screen(Has touch screen or not), and wifi (Has wifi or not)

Related Findings

"Exploring the Factors That Influence Consumer's Purchase of Mobile Phones" is a publication in Researchgate website. The main goal of this paper is to identify what factors influence and mobile phones. The authors think the brand loyalty, price, quality, social influence and mobile features are the factors that influence consumers to purchase mobile phones. Brand loyalty is most important factor than other factors which means the consumers are willing to pay a premium price for the brand.

This publication's conclusion is consistent with my current analysis of the data that the brand has a great influence on the price of mobile phones. During my analysis of the dataset, I find that most mobile phone price change depending on the specifications of the phone, but there are many special examples. Some phones' specifications not as good as other phones, but the price is not lower than other phones.

Similar Project

There are a lot of projects like my project. I found a project called mobile phone price prediction wrote by Vikramadiytya Singh Bhati. This project is very similar to my project. We use the dataset, have same analysis steps and same results so far. However, the analysis methods used by him and me are different, and the future models will also be different, so the final outcome may be different.

Exploratory Data Analysis

At first step, I use the jupter notebook to read the dataset to see all features in the data. I found out that the data looks really good, which means I don't need to do any transformations and cleaning. Then, I visualized the relationship between mobile specifications and price range to determine which specification has great impact on the price. I found that there are two specifications that have a greater impact on the price, the others are not. As shown in the graphs below, the battery power and random-access memory in megabytes are important features that higher battery power or higher random-access memory in megabytes tend to have higher price.

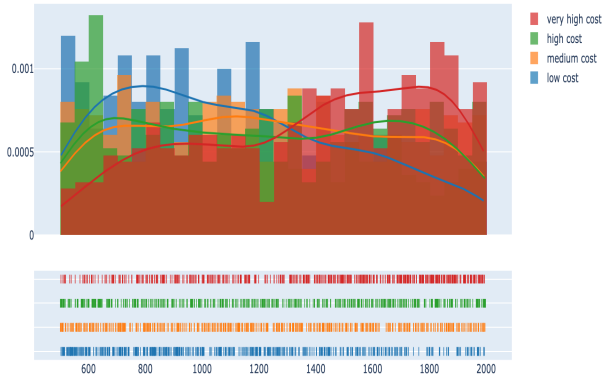

Figure 1. battery power vs price

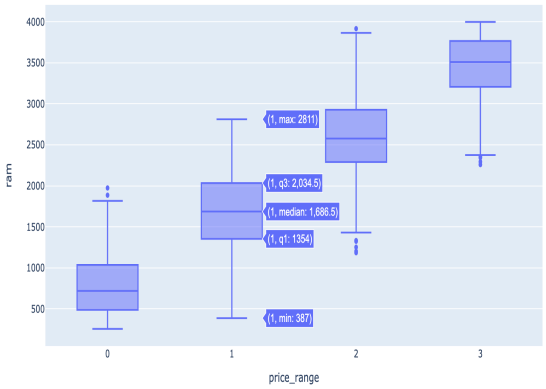

Figure 2. random-access memory vs price

Data Preprocessing

Before I do the modeling for my project, I have to complete the data preprocessing. First, I need to complete the data preprocessing to prepare for modeling, because each feature has a different range. I need Standardize features by removing the mean and scaling to unit variance. "standardScale" makes the processed data conform to the standard normal distribution, ie the mean is 0 and the standard deviation is 1. This can speed up the speed of gradient descent to find the

optimal solution and improve the accuracy for the Non-probabilistic model. Secondly, I divided the data into a training set and a test set. 80% of the data is used as the training set, and 20% of the data is used as the test set.

Models

Nature of the Problem

As Output labels shows, 0 means low cost, 1 means medium cost, 2 means high cost and 3 means very high cost. We can be sure that this is a multiclass classification problem. A sample belongs to and belongs to only one of multiple classes, a sample can only belong to one class, and different classes are mutually exclusive. Based on the nature of the problem, I chose Logistic Regression, Neural Network and Random Forest to predict this problem. For machine learning models, I used the sklearn package. Because this is a multi-classification problem, I used Mean Accuracy as the metric to evaluate the model.

Logistic Regression

I firstly use Logistic Regression. Logistic Regression is mainly used for binary classification. The algorithm uses the one-vs-rest (OvR) scheme when facing multiclass case. One-Vs-All‘s idea is to turn a multi-class problem into multiple two-class problems. The idea of the change is just as the method name describes, choose one of the categories to be positive and make all other categories to be negative. Its advantages are high efficiency and fast speed, and it can train as many classifiers as there are categories. But its shortcomings are also obvious. Logistic regression belongs to the generalized linear model, so it does not perform well on linear inseparable problems. Then, I use the confusion matrix to show the results. Confusion matrix is a situation analysis table that summarizes the prediction results of the classification model in machine learning. The records in the data set are summarized in a matrix form according to two criteria: the True Labels and the predicted labels by the classification model. Where the rows of the matrix represent the true values and the columns of the matrix represent the predicted values; From the confusion matrix, I found that the prediction error only occurred in adjacent labels. In other words, the model never predicted high cost as low cost.

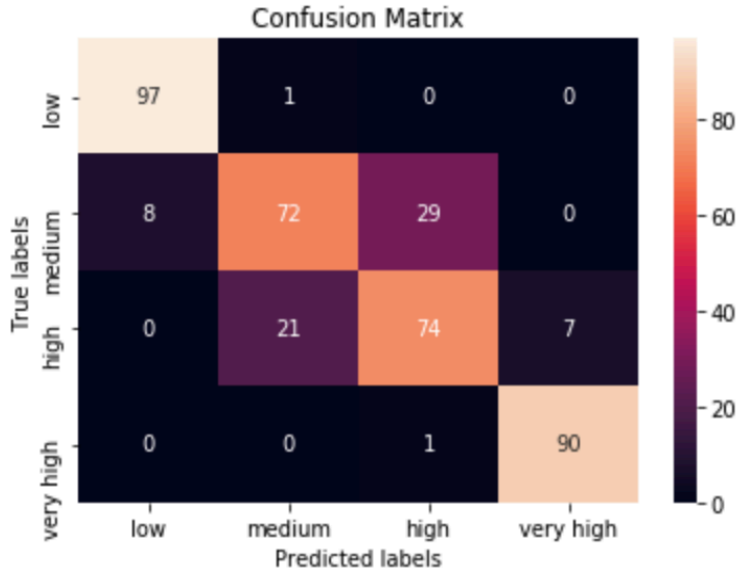

Figure 3

Neural Network

Next, I used a neural network and wanted to obtain higher accuracy through its powerful non-linear capabilities. Neural Network has the Ability to learn and build models of nonlinear complex relationships. In addition, the neural network can be generalized. After learning from the initial input and its relationship, it can also infer the unknown relationship from the unknown data, so that the model can generalize and predict the unknown data. The disadvantage is that the neural network requires a lot of training data. After hyperparameter tuning, I chose Relu as Activation function and Adam as Optimization Algorithm. From the confusion matrix that the neural network performs better than logistic regression, especially in medium and high cost.

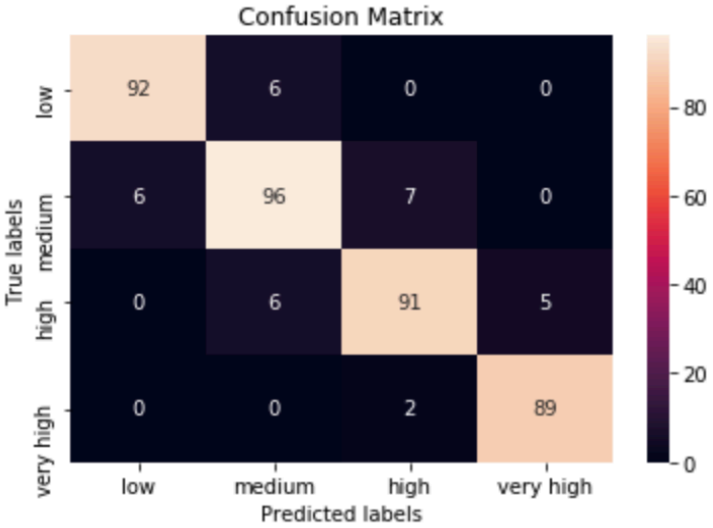

Figure 4

Random Forest

Lastly, I choose to use random forest. Random forests can handle high-dimensional data without dimensionality reduction or feature selection. In addition, it is not easy to overfit. The most important thing is that it can judge the importance of features. The disadvantage is that it does not perform well in noisy data sets. From the confusion matrix that the random forest performance is better than logistic regression but worse than neural network.

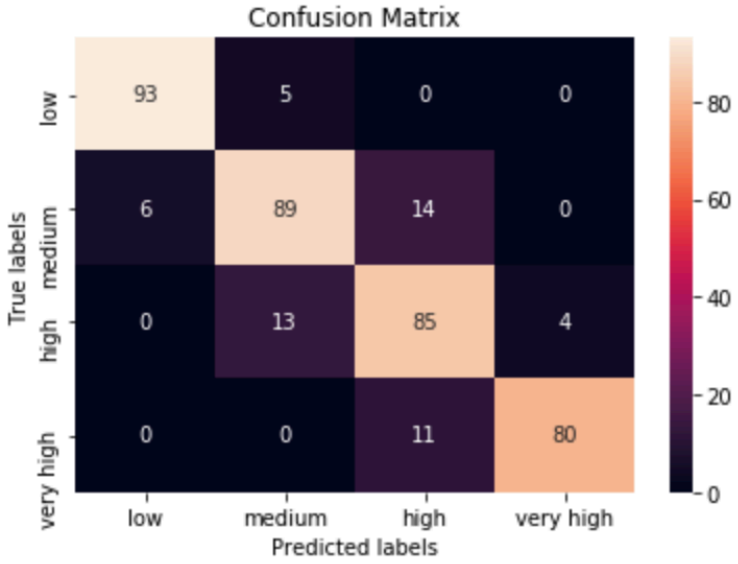

Figure 5

Accuracy

The accuracy of the logistic regression model in the training set is 86.75%, and the mean accuracy in the test set is 83.25%.

The accuracy of neural network has been improved. The accuracy of the training set reached 100% because of the powerful memory of the neural network. In the test set, accuracy is as high as 92%.

The Accuracy of the random forest the Train Data is 98.75% and Mean Accuracy for the Test Data is 86.75%. Its performance is better than logistic regression but worse than neural network.

Results

At EDA part, I performed distribution analysis and we were only watching the impact of single feature. Therefore, I use machine learning to watch the joint influence. After, I applied logistic regression, neural network and random forest, we can clearly see that Ram is the most important feature, and its importance is much higher than other features that are given by Random forests. Battery power is another important feature. As shown in below, from the plot, mobile feature like 3g, 4g, wifi, Bluetooth doesn't influence price that much. And this result is same with my exception, because from my life experience, both cheap and expensive mobile phones have these configurations.

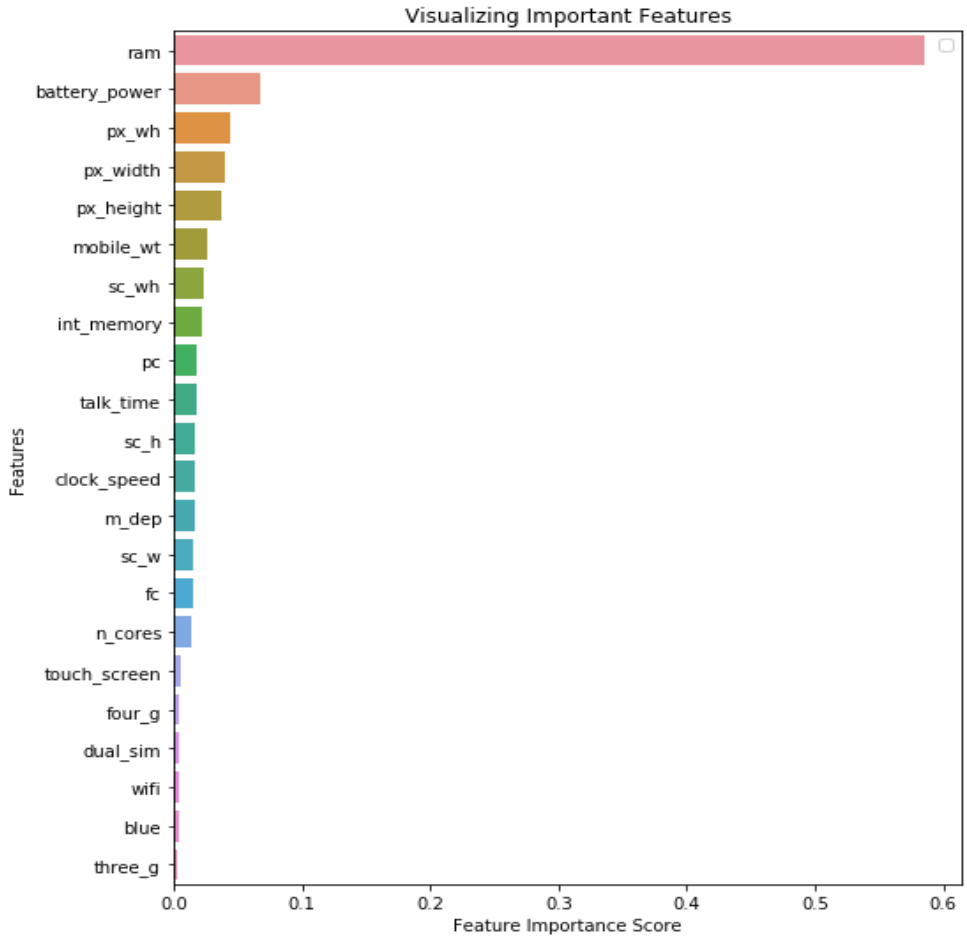

Figure 6

Conclusion

One problem of this project is that the price data we obtained is not a definite value range. In other words, my model didn't tell us a phone should cost something between a specific price range like between $900 and $950. My project uses four categories to replace specific price and the four categories represent very low cost, medium cost, high cost and very high cost. I think use four categories can make our data and results more useable. As we know, the price of the mobile is changing ever year. The mobile phone we bought in 2010 with $700 must be a completely different configuration from the mobile we bought with $700 now, for example, the iPhone 4 sells for $699 in 2010 with 32 GB, and iPhone 8 sells for $699 with 68 GB now. Even with the same configuration, the price in 2010 and the price in 2020 will be very different. The second reason why I chose to use four categories is Consumer consumption levels are constantly changing. Therefore, in 2010, the $700 mobile phone was very high cost, but now it can only be considered high cost. When I use four categories to instead specific price range, the mobile phone manufactures can use my model for several years. They use my model to decide which price range their mobile phones belong to base on their mobile phone features, whether it is low or high. Then

determine the price based on the current value of the currency and the market. Compare three models' accuracy, we can know the Neural Network is the best model for prediction. Therefore, I suggest mobile phone manufacture use neural network model to do the prediction.

Work Extension

For this project, I decided to propose a method to simulate a specific price range to help us understand what low cost is, what is medium cost, what is high cost and what is very high cost. I use Wikipedia as my resources to collect the data with specific price of the mobile phone. I Randomly select 16 phones for the list and google the release price of that year and the RAM (in megabytes). From the results I got from the machine learning and EAD WE Clearly K now the Ram are the most important feature, so I chose the ram as the key feature. I retrieved the Data of the Year 2015 and 2016 as shown in figure 7 and 8.

	Name	RAM	price
0	Acer Liquid Z630	2000	216
1	Alcatel One Touch Idol 3	2000	250
2	Droid Turbo 2	3000	520
3	Honor 5X	2000	200
4	HTC Butterfly 3	3000	530
5	iPhone 6S	2000	550
6	Lava Pixel V1	2000	150
7	Lenovo A6000	1000	150
8	LG G4	3000	550
9	Microsoft Lumia 430	1000	120

Figure 7

	Name	RAM	price
0	Alcatel Idol 4	3000	377
1	BlackBerry DTEK50	3000	300
2	Cat S60	3000	600
3	HP Elite x3	4000	700
4	iPhone SE	2000	399
5	Pixel	4000	650
6	Samsung Galaxy A8	3000	800
7	Sony Xperia X	3000	645
8	LG G5	4000	700
9	Redmi 3	2000	150

Figure 8

After we scatter plot the actual price of the mobile phone and the memory of the mobile phone, we found that the price and the memory have a strong positive correlation.

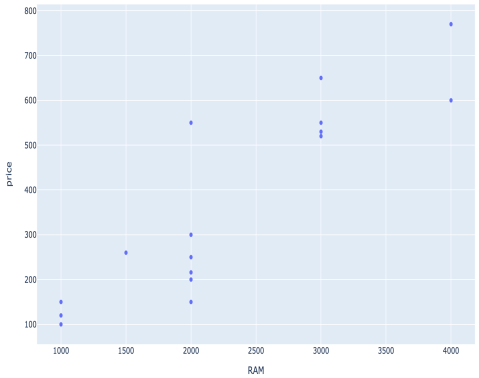

Figure 9

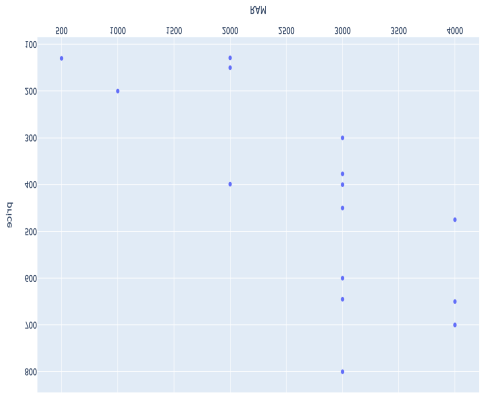

Figure 10

This is same as the trend we found in training data. In addition, you can notice the points have tendency to move to the right between year 2015 and 2016. This is also logical. With the

development of time and technology, larger memory is a trend. This is more obvious in Box plot. Compare with the box plot of ram and price range in training data, they are very similar.

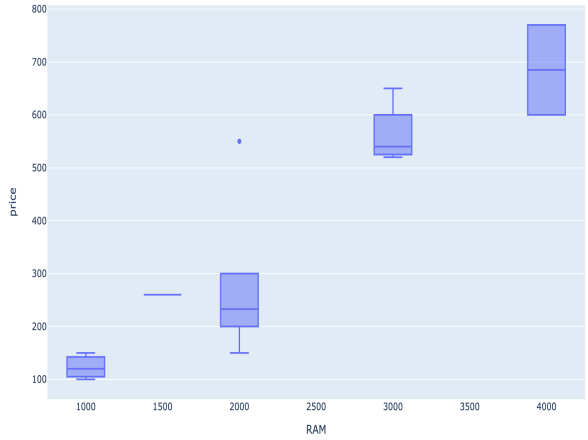

Figure 11

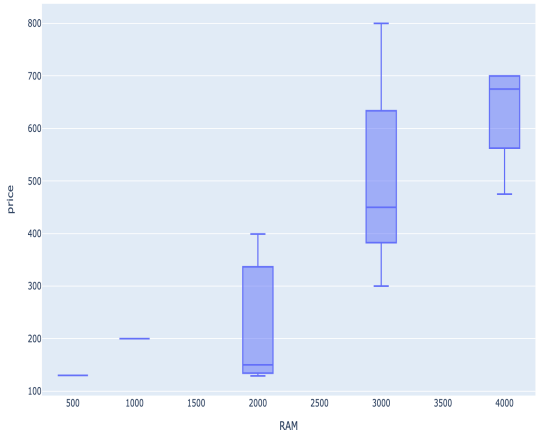

Figure 12

It can easily tell the price range for each price tag by looking at the box plot. More specifically, we can use the first quartile (Q1) and third quartile(Q3) to decide to decide the demarcation point of the price range.

From the result, we can know the same price tag may have different range in different years. Such as in 2015 the median price is 200-500, and in 2016 the median price change to 200-350.

Price Tag	**2015**	**2016**
Low	<200	<200
Median	200 ~ 500	200 ~ 350
High	500 ~ 600	350 ~ 600
Very High	600<	600<

During my work, I have other findings that related to my project. my models are less than 100% accuracy Under actual conditions because Price can also be affected by Brand and Country. For example, the Freedom 251 is a smartphone that was initially offered for sale in India at the promotional price of ₹251 (the equivalent of $3.54 as of 2020). It was sold by Ringing Bells Private Limited and was marketed as the world's cheapest smartphone.

For another example, I compare two phones are Xiaomi mi 2 and iPhone 5. Both phones were released in the same year, and iPhone 5 sold for $650, the Xiaomi Mi2 sold for $300. As we know the ram is the most important feature and will has a big impact of the price that higher ram lead to higher price. However, the mi 2 which has higher ram sold for lower price. These two special examples fully prove that the country and the brand will affect the price of the mobile phone.

Future Work

Use Web scraping technique to obtain the data with other important features, like brand, country, new design to make my models more accuracy under actual conditions.

Reference:

Abhishek Sharma. (2017). Mobile Price Classification. Kaggle. Retrieved from https://www.kaggle.com/iabhishekofficial/mobile-price-classification#test.csv

Vikramaditya SIngh Bhati. (2017). Mobile Phone Price Prediction. Retrieved from **https://www.kaggle.com/vikramb/mobile-price-prediction**

Mohd Yusuf, Bibi Noraini & Hock, Lim & Abd Rashid, Intan & sa'aban, Syahira & Abdullah, Muhammad. (2015). Exploring the Factors That Influence Consumer's Purchase of Mobile Phones. Journal of Advance Research in Business, Management and Accounting.