Smith School of Business, Queens University | Master of Management Analytics Toronto

MMA 867
Predictive Modelling

Professor Jue Wang

Project Report
11th June 2022 | 11:59 PM

Team Adelaide

Student Name	Student Number
Ishaan Sinha	20290167
Jiali Bai	20316342
Juefei Chen	20311141
Kripa Shanker Nayak	20315710
Oluwaseun Adelana	20310041
Yichen Yuan	20315046

Order of files:

Filename	Pages	Comments and/or Instructions
MMA867_ProjectReport	8	

Additional Comments:

Mobile Price Classification

EXECUTIVE SUMMARY

Team Adelaide, a new entrant in the smartphone industry, has recently completed its product development phase. Our product line has different types of smartphones, such as touch screen vs. keypad phones, dual sim vs. single sim phones, etc. Having developed all these new smartphones, we now want to classify them into different price segments. People have varying uses for smartphones, ranging from texting and checking emails to photography and even banking. Thus, there is a demand for smartphones at all price points, and we have to offer the best mix of features at all these price points to attract customers effectively in a highly competitive market.

To achieve this, we collected current data for the current smartphone market. After adequately preparing and analyzing our data, we built a classification model using logistic regression to classify smartphones according to their retail prices. We found the most sought-after features at each price range from this model and can now build a pricing strategy around this information.

Introduction

The competition in the smartphone industry has always been intensive; it is still increasing and is projected to grow till 2030. According to the latest study from Statista, the Global Smartphone market was valued at USD 273.9 Billion in 2021 and is expected to reach USD 520.7 Billion by 2030. The top 3 industry giants, Samsung, Apple, and Xiaomi, share more than 53% of the market share. These companies offer a large variety of smartphones with different features at various price ranges, and the competition in the industry occurs within these price ranges. Using iPhone as an example, to accommodate different customer groups, each generation of iPhone may have at least three models, such as iPhone starting at $899, iPhone Pro starting at $999, or iPhone Pro Max starting at $1099. Their appearance might be identical, but they are designed with different features such as different processing chips, camera resolution, and battery life. Thus, offering the most attractive mix of features at each price segment is essential.

Despite the brutal competition and rapid technology development, we plan to start our own mobile business. After years of product development planning, it has come to the stage that we need to estimate the price of our self-developed mobile devices. We offer a wide variety of offerings with a different combination of features. We need to develop an effective pricing strategy so that these can be put into an accurate, rational segment.

Data Collection and Description

We did secondary research to collect data for our model-building process. Data points for various cell phone attributes were collected through publicly available information for current mobile manufacturers. Variables associated with the features offered for each product were closely analyzed, and finally, 21 variables were selected, including price category. We collected market data for all 21 variables based on multiple features such as battery power, Bluetooth, screen size, pixel density, processing power, dual sim capability, network options (3G or 4G), etc. We also bucketed our products into four price categories: Low Cost, Medium Cost, High Cost, and Very High Cost, and collected equal observations for each price point. A total of 2000 observations were collected with 14 numerical and 7 categorical variables for our model building process. In our model, Price is the response variable, whereas other variables are considered predictor variables.

We further refined and analyzed our data by performing various data cleaning and exploration techniques. We checked our data for any missing values, and in scenarios where missing data were found, we used the missing value imputation technique to replace null values. We also checked the dataset for duplicated columns to ensure data cleanliness. After data cleaning, we identified all the categorical variables within our dataset, such as Bluetooth, 3G, 4G, etc., and transformed them into dummy variables. Post data transformation, we analyzed the distribution of all the variables within the dataset to understand their pattern and frequency. While observing the distributions, we found variables such as battery power, mobile weight, RAM, primary camera megapixels, etc., to be approximately uniformly distributed, whereas variables such as front camera megapixels, pixel height, screen width, etc. had a left-skewed distribution. We also created a correlation matrix between variables to check

for multicollinearity. As per the correlation matrix, we observed that variables such as Primary camera, front camera, pixel width, pixel height, 3G, 4G, price range, and RAM were highly correlated. Once our data was cleansed and analyzed, we started our model-building process.

Model Development

We started our model-building process by applying the multinomial logistic regression model as our base model. The dependent variable was price range, representing different price levels, and the rest of the variables were independent variables. 70% of the data were randomly allocated as the training set and the remaining 30% as our testing set. For the base model, we ran the regression without any transformations of independent variables, and the accuracy of the base model was 65%.

We then applied feature scaling to the model to bring all the features down to a standard scale. Before applying feature scaling, we noticed that some features have a wider spread of ranges than others. For example, the data spread for battery power ranged from 600 to 2000, while the data spread for mobile depth only ranged from 0 to 1. We applied standardization to our model to avoid such features with wider ranges dominating the distance metrics. Through this technique, the means for all the features were moved to zero, and distances were scaled to unit standard deviation, indicated by a z-score. Therefore, the impact from each feature was scaled to a standard range for our model. After feature scaling, the accuracy of the model increased to 98%.

During our model building process, we ran the variance inflation factor (VIF) to measure multicollinearity in our model. We noticed that some variables, such as mobile weight, and screen pixel width showed high VIF scores in the model. Therefore, we decided to use principal component analysis to eliminate the impact of highly correlated variables. We created new components in which variables are grouped together within components that have their own weight. The components created are also proportional to the model; for instance, PC4, the one with the highest proportion, weighted over 8% of the model. In our final model, we used the first 18 components with more significant variances than the rest, representing over 96% of

the cumulative proportion. The result gave the highest accuracy by choosing these 18 components in our model.

Finally, we did L2 regularization for feature selection. After applying principal component analysis and L2 regularization, the accuracy improved to 98%.

Model Output Interpretation

In evaluating the multinominal logistic regression model, we had 98.7% accuracy on the train set and 97.8% on the test set. The train and test scores performed very well and are very close to each other, implying the model was not likely under-fitted or overfitted. According to the confusion matrix (figure 1) and classification report (figure 2), this model's precision, recall, and f1 scores also performed exceptionally well for each category. The precision ranged from 0.97 to 0.99, the recall ranged from 0.96 to 0.99, and the f1 score ranged from 0.97 to 0.99 for each category. In the test set, the model correctly classified 147, 144, 148, and 148 observations of the low, medium, high, and very high prices. To sum up, the model has done an excellent job predicting all the classes.

First, we looked at the coefficient chart (figure 3) to interpret the model results. The chart showed that high values for PC 6 increase the likelihood that the mobile is more expensive, which is the opposite of PC 14, high values for PC 14 increase the probability that the mobile is cheaper. Therefore, we dug deeper into these two components. According to the heatmap of the principal component loadings (figure 4), the PC 6 was positively correlated with features such as Battery power, internal memory, and Ram. On the other hand, PC 14 was positively correlated with features such as mobile weight and talk time without charging which meant those features had less impact on increasing the mobile phone’s price. In conclusion, customers tended to spend more money on phones with considerable battery power, internal memory, and RAM.

Key Outcomes and Future Steps

Some of the key outcomes and potential benefits identified for the model developed include:

- Collect and process extensive data from different sources to predict the market value of mobile devices. The model helps understand how our product fits in the value chain and whether it can be priced as a low, medium, or high product category.
- Identify what potential customers may consider value for their money based on the comparable features of our mobile device with other products in the market.
- Inform the development of an optimized pricing strategy – set a price that not only covers the cost of production but generates a substantial return, avoids missed customers, gains the right customers, and offers customers a price that portrays a good bargain and value for money.

As the next steps for our project, we plan on building a time series model to predict the performance of different features in the market.

Appendix

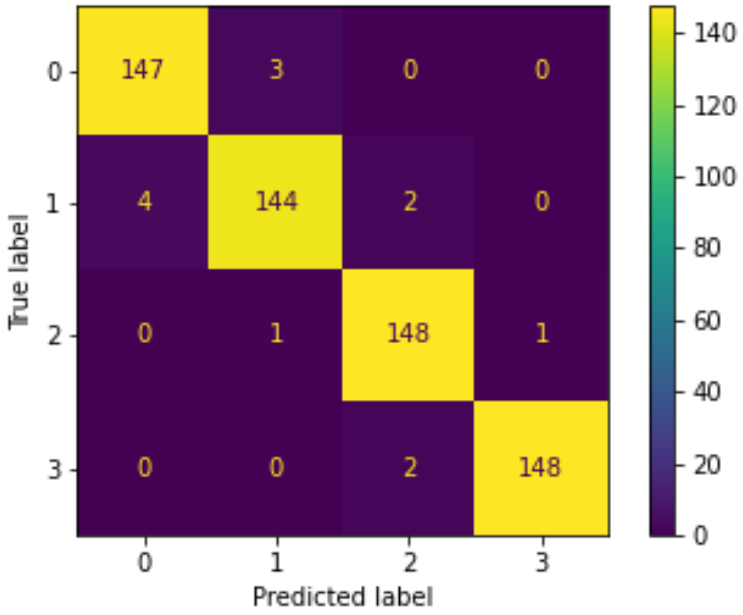

Figure 1

	precision	recall	f1-score	support
0	0.97	0.98	0.98	150
1	0.97	0.96	0.97	150
2	0.97	0.99	0.98	150
3	0.99	0.99	0.99	150
accuracy			0.98	600
macro avg	0.98	0.98	0.98	600
weighted avg	0.98	0.98	0.98	600

Figure 2

Component	Low (0)	Medium (1)	High (2)	Very High (3)
PC01	2.56	0.74	-0.74	-2.57
PC02	7.09	2.38	-2.14	-7.34
PC03	-7.06	-2.34	2.21	7.19
PC04	-1.29	-0.42	0.57	1.14
PC05	-5.34	-1.66	1.71	5.29
PC06	-24.05	-7.70	7.86	23.89
PC07	2.60	0.74	-0.84	-2.50
PC08	5.69	1.85	-1.80	-5.75
PC09	-2.25	-0.75	0.82	2.19
PC10	21.49	7.03	-7.09	-21.44
PC11	6.10	2.04	-2.16	-5.98
PC12	18.73	6.06	-6.06	-18.73
PC13	-1.12	-0.25	0.46	0.91
PC14	22.60	7.31	-7.28	-22.64
PC15	-8.64	-2.68	2.70	8.61
PC16	-15.61	-4.84	5.24	15.21
PC17	-2.99	-0.68	0.67	3.01
PC18	0.30	-0.15	0.08	-0.23

Figure 3

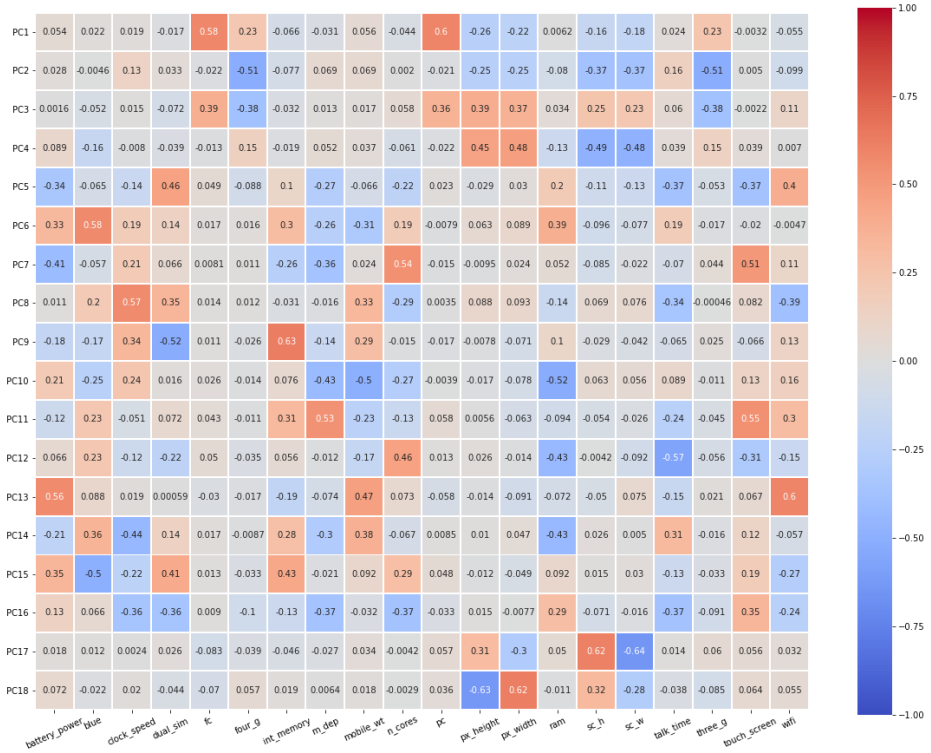

Figure 4