Big Data Project

Apps from the Apple iOS

ECON 695 Zishan Bai

Abstract

When we download App from App Store, considering the rating of this App will be a common way to decide whether it is a good App. This project is aiming at which factors that influence the rating of an Apps a lot and whether an unfree Apps will have higher ratings than free Apps.

There are three parts in this project, the first part will describe the dataset that will be used in this project including its observations, its variables, and the summary of the data. By plotting different variables against the ratings, it will give us some intuitions about which factors are related to the ratings of an App. The second is my method to choose the best model for this dataset and the evaluation of the model. Finally, I will use my model to do some analysis and give the conclusion in the last part.

I. Data

i. Data Describing

Firstly, the data in this big data project is mobile app statistics from the Apple iOS apple store, which has 7197 observations, and each observation has 16 variables. There are two different ratings in this dataset. The one is the user rating of each App for all versions (user_rating) and the other one is the use rating of each App for the current version(user_rating_ver). In order to decide the updated condition of each App, I decided to use the rating of each App for current versions as the dependent variable. Figure 1 in Appendix is the summary of the dataset. From the summary, we know that the mean of the user rating for current version is 3.25 and 75% of the Apps have the ratings below 4.5. Figure 2 is the histogram of user rating. We can see that most App have a rating of 4.5. The number of Apps with rating between 0.5 to 3 is small but there still have around 1500 Apps which have a rating of 0.5.

ii. Variables Selection

Refer to Figure 1, some variables are the description of each App such as track_name or prime_genre. Therefore, I did not plan to use these variables since they are not useful to account for a quantitative analysis. By plotting some plots between variables, I plan to find some necessary variables. Adding that not every variable is useful in this dataset to influence the rating of an App, I will select some variables that will be informative and influence the rating of an App intuitionally. I select four variables listed below as the independent variables.

price: App Price amount
size_bytes: App Size in Bytes
sup_devices.num: Number of supporting devices
lang.num: Number of supported languages

II. Analysis

i. <u>Explore single relationship</u>

Firstly, I select the independent variables I choose from the origin data and generate a new dataset. Next, I plan to explore the relationship of each variable with the ratings of current version, which gives us an idea about how each variable influences the rating of an APP.

Price & Rating

In intuition, people prefer to download an application with the rating of 4 in a 5-scale system. Therefore, in this project the rating above 4 (include) will be regarded as high-rating group, and ratings below 4 will be low-rating group. From the result, I find that 4406 Apps have a rating above 4, which means that around 61% of Apps in the store have a high rating. Next, I divide price as free group and unfree group. From the result that 4056 Apps are free and 3141 Apps are not free, which means that more than half applications are free. Figure 3 shows the relationship between price and ratings, indicating that the number of high rating Apps in free group is more than that in unfree group.

Size & Rating

By observing the data of App Size in Bytes, I found that this data is right skewed. Therefore, I tried to use the log transformation here so that I can compare the distribution between the ratings and App size. Figure 4 shows the result of these two. 18.2 is the standard. The size below 18.2 has more Apps with rating lower than 4 while the size over 18.2 has more APPs with rating above 4.

Number of supporting devices & Rating

From the summary, we know that the medium of sup_devices.num is 37, so I divide the data as above 37 devices and under that. Applying the same way as above, Figure 5 shows the relationship between ratings and number of supporting devices. It is interesting to see that most application supporting more than 37 devices, but there are still a lot of applications with ratings lower than 4 even though they can support more than 37 devices.

Number of supported languages & Rating

Applied the same way, Figure 6 use 5 as the dividing line of number of supported languages. The differences between ratings are distinctive if the number of supported languages devices is over 5, and it doesn't influence the ratings a lot in less than 5 languages.

ii. Modelling

As for binary logistic regression, the number of dependent variables is two, whereas the number of dependent variables for multinomial logistic regression is more than two. Since in this project we only consider high rating group and low rating group, thus I plan to use *Binary Logistic Regression* to predict the rating of each APP, and evaluate the accuracy of my model.

Since all variables in this model are categorical variable except log transformation of App size in Bytes, there is no worry about multicollinearity. The model is as follows:

$$\text{logit}(p) = \beta_0 + \beta_1*\text{price} + \beta_2*\text{ sup\_devices.num} + \beta_3*\text{ lang.num} + \beta_4*\text{logsize}$$

a). *Result of the Model*

Recall that user_rating = 1 means high-rating group, price = 1 means paid App, sup_devices.num = 1 means supporting over 37 devices, and lang.num = 1 means supporting over 5 languages, the model regards user_rating = 0 as the default reference category. Figure 7 in the Appendix is the summary of the regression. The original model gives us log odds of dependent variables versus the reference category. Therefore, I extract the coefficients from the model and exponentiate them to get the real estimate. The result is showed in Figure 8. Refer to Figure 7, all of p-values are very close to 0, so we can say that the estimates in the model are all statistically significant.

b). *Model evaluation*

First, I check the model fit information. To compare with the current model, I run an only-intercept model and compare the change of Residual deviance. Figure 9 shows the result of only-intercept model. We can see that the Residual deviance has changed from 9613.7 to 9010.8, which is a decrease of 602.9. This decrease suggests that the current explains a significant amount of the original variability.

Second, I check the accuracy of the model by calculating prediction accuracy. Figure 10 shows the result. As the result, the percentage of correct prediction for high-rating group is 84.1%, and the total percentage of correct prediction is 66.3%. In the next part, this project will talk about the interpretation of each coefficient in this model.

c). *Interpretation of Coefficients*

In Figure 7, all of coefficients estimates are bigger than zero, meaning that the log-odds are all positive and these variables did influence the rating of an application. The following is the interpretation of each coefficient according to Figure 8.

Price: Holding all others variables as constants, the odds of getting high ratings for paid Apps over the odds of getting high ratings for free Apps is 1.3767571, meaning that the odds of getting high ratings for paid Apps are 37.7% higher than that of free Apps.

Number of supporting devices: Holding all others variables as constants, the odds of getting high ratings for Apps supporting more than 37 devices over the odds of getting high ratings for Apps supporting no more than 37 devices is 1.2560670, meaning that the odds of over 37 supporting Apps are 25.6% higher than that of below 37 supporting Apps.

Number of supported languages: Holding all others variables as constants, the odds of getting high ratings for Apps supporting more than 5 languages over the odds of getting high ratings for Apps supporting no more than 5 languages is 2.7125077, meaning that the odds of over 37 supporting Apps are 171.2% higher than that of below 37 supporting Apps.

Size: Since the log size here has been standardized, I interpreted it by looking for one standard deviation above the average, and we can see that the odds of getting high ratings is 42.2% higher if one standard deviation above.

III. Discussion

I also tried the multinomial logistic regression using medium rating as the reference category (see Figure 11). However, AIC of multinomial regression is 14629.55, which is much higher than the binary logistic regression, and it doesn't give me more valid explanation for my analysis. Therefore, I still think binary logistic regression is more useful.

IV. Conclusion

In conclusion, unfree Apps have more chance to be a good rating App, so it is possible that paid products have better quality in the same category. In particular, the influence of supporting language is the strongest one, indicating that if a company want to improve the rating of their products, one efficient way is to serve for different languages as much as possible.

Appendix

Figure 1: Summary of Data

```
       X              id             track_name          size_bytes            currency
 Min.   :    1   Min.   :2.817e+08   Length:7197        Min.   :5.898e+05   Length:7197
 1st Qu.: 2090   1st Qu.:6.001e+08   Class :character   1st Qu.:4.692e+07   Class :character
 Median : 4380   Median :9.781e+08   Mode  :character   Median :9.715e+07   Mode  :character
 Mean   : 4759   Mean   :8.631e+08                      Mean   :1.991e+08
 3rd Qu.: 7223   3rd Qu.:1.082e+09                      3rd Qu.:1.819e+08
 Max.   :11097   Max.   :1.188e+09                      Max.   :4.026e+09
     price          rating_count_tot   rating_count_ver     user_rating     user_rating_ver
 Min.   :  0.000   Min.   :      0   Min.   :     0.0   Min.   :0.000   Min.   :0.000
 1st Qu.:  0.000   1st Qu.:     28   1st Qu.:     1.0   1st Qu.:3.500   1st Qu.:2.500
 Median :  0.000   Median :    300   Median :    23.0   Median :4.000   Median :4.000
 Mean   :  1.726   Mean   :  12893   Mean   :   460.4   Mean   :3.527   Mean   :3.254
 3rd Qu.:  1.990   3rd Qu.:   2793   3rd Qu.:   140.0   3rd Qu.:4.500   3rd Qu.:4.500
 Max.   :299.990   Max.   :2974676   Max.   :177050.0   Max.   :5.000   Max.   :5.000
     ver             cont_rating         prime_genre         sup_devices.num ipadSc_urls.num
 Length:7197        Length:7197        Length:7197          Min.   : 9.00   Min.   :0.000
 Class :character   Class :character   Class :character     1st Qu.:37.00   1st Qu.:3.000
 Mode  :character   Mode  :character   Mode  :character     Median :37.00   Median :5.000
                                                            Mean   :37.36   Mean   :3.707
                                                            3rd Qu.:38.00   3rd Qu.:5.000
                                                            Max.   :47.00   Max.   :5.000
    lang.num          vpp_lic
 Min.   : 0.000   Min.   :0.0000
 1st Qu.: 1.000   1st Qu.:1.0000
 Median : 1.000   Median :1.0000
 Mean   : 5.435   Mean   :0.9931
 3rd Qu.: 8.000   3rd Qu.:1.0000
 Max.   :75.000   Max.   :1.0000
```

Figure 2: Histogram of Ratings

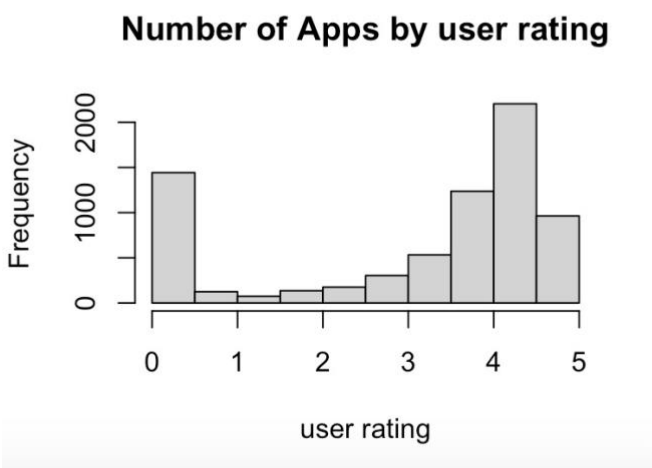

Figure 3: Price & Ratings

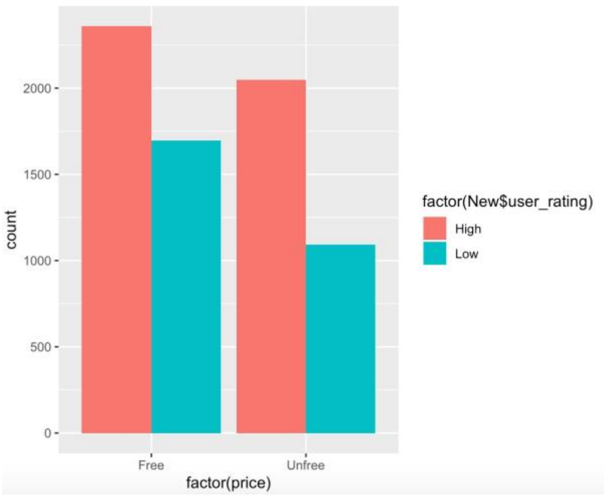

Figure 4: Size & Ratings

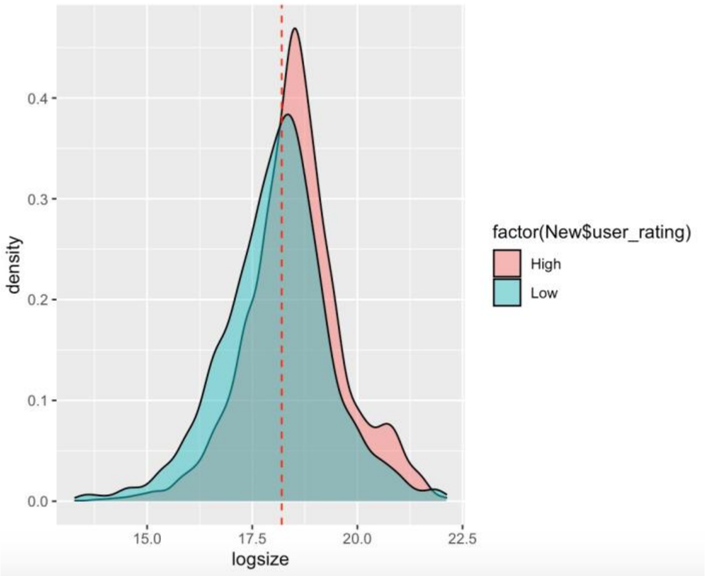

Figure 5: Number of supporting & Ratings

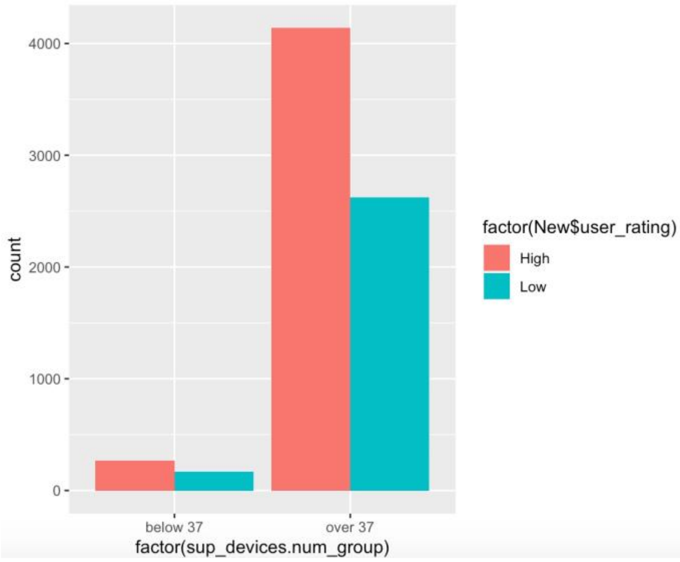

Figure 6: Language & Rating

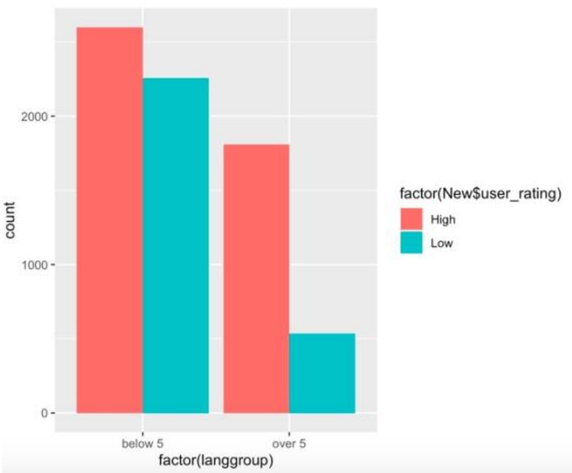

Figure 7: Summary of Logistic Regression

```
Call:
glm(formula = user_rating ~ price + sup_devices.num + lang.num +
    logsize, family = "binomial", data = New)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2310  -1.1791   0.7062   1.0315   1.7584

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     -0.17158    0.11010  -1.558   0.1191
price            0.31973    0.05153   6.205 5.49e-10 ***
sup_devices.num  0.22799    0.10790   2.113   0.0346 *
lang.num         0.99787    0.05790  17.235  < 2e-16 ***
logsize          0.35238    0.02654  13.278  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 9611.7  on 7196  degrees of freedom
Residual deviance: 9000.8  on 7192  degrees of freedom
AIC: 9010.8

Number of Fisher Scoring iterations: 4
```

Figure 8: Log Transformation of Estimate

```
(Intercept)           price sup_devices.num        lang.num
  0.8423354       1.3767571       1.2560670       2.7125077
    logsize
  1.4224516
```

Figure 9: Only-Intercept Regression

```
Call:
glm(formula = user_rating ~ 1, family = "binomial", data = New)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3764  -1.3764   0.9907   0.9907   0.9907

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.45657    0.02419   18.87   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 9611.7  on 7196  degrees of freedom
Residual deviance: 9611.7  on 7196  degrees of freedom
AIC: 9613.7

Number of Fisher Scoring iterations: 4
```

Figure 10:

```
       predir
        High  Low
  High  3706  700
  Low   1807  984
```

Figure 11: Multinomial Logistic Regression

```
Call:
multinom(formula = rating2 ~ price + sup_devices.num + lang.num +
    logsize, data = New, model = TRUE)

Coefficients:
            (Intercept)     price2 sup_devices.num2 lang.num2     logsize
low rating   -1.5374134 0.70445769       -0.5414062  1.120253 -0.37739120
high rating   0.5033101 0.09472869       -0.3517666 -0.308716  0.04705706

Std. Errors:
            (Intercept)     price2 sup_devices.num2  lang.num2    logsize
low rating   0.08693563 0.06817423        0.1518525 0.08354194 0.03438930
high rating  0.05553262 0.05575249        0.1097817 0.05750916 0.02834334

Residual Deviance: 14609.55
AIC: 14629.55
```