Stat 403 Final Project Report

Group II: IOS Mobile Apps

1 Introduction

During recent years, mobile apps are experiencing a booming period due to the increasing use of phones in people's daily lives. Mobile app users have found themselves more and more difficult in selecting a satisfying app within the category, especially when there are multiple similar apps at the same time. While people tend to select apps with higher user rating score, several questions are raised during the exploration of the dataset: (a) what are the primary factors that contribute to the rating score of an app? And (b) how could we predict the rating score of an app using the selected features?

The dataset used in this report was obtained from Kaggle (https://www.kaggle.com/ramamet4/app-store-apple-data-set-10k-apps), which was originally extracted from the iTunes Search API at the Apple Inc website. The original dataset, which was collected in July 2017, contains details of 7197 apps in the IOS system with 16 variables in total (fig.1).

Figure 1: Names for 16 variables contained in the dataset. The column names in the dataset were shown in bold and followed by their discriscription.

In this study, the average user rating (user_rating) was the target of our research and was set as the dependent variable. The aim of this report is to (a)

use bootstrap, an effective resampling tool, to provide a good estimation of the average user rating based on different characteristics of the app within the genre; and (b) identify the important factors that contribute to the average user rating and build regression models for predicting the user rating.

2 Methods

Processing Data (Exploratory Data Analysis)
In the Apple store, the lowest score of the user rating that an app can have is 1. Therefore, the dataset was adjusted by filtering out 929 apps with 0 rating counts in total since it provides insufficient information for our data analysis.

Among the dataset, app ID, app name and vpp_lic are not relevant to the data analysis. On the other hand, all apps were collected from the Apple store in the U.S. resulting in the currency all in USD. Thus, for this study, eight variables including app name (track_name), app id (Id), currency (currency), version (ver), current average user rating (user_rating_ver), current version's user rating count (rating_count_ver), content rating for different age groups (cont_rating) and enabled vpp device based licensing (vpp_lic) were removed from the dataset because they either contribute little or are weakly related to the average user rating according to the correlation heat map (fig.2).

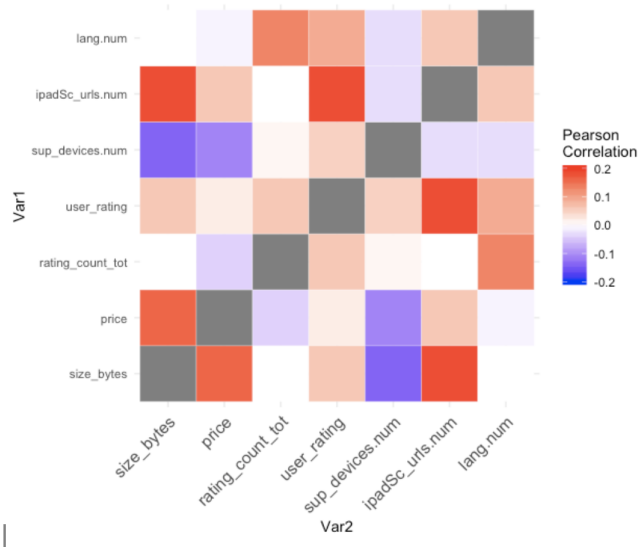

Figure 2: The Correlation Heat Map for Variables

The selected continuous variables are user rating (user_rating), size in

bytes (size_bytes) and price (price). The discrete variables are total rating count (rating_count_tot), screenshot in iPad devices (ipadSc_urls.num), number of supporting languages (lang.num) and number of supporting devices (sup_devices.num).

The correlation heat map shows that most variables are correlated with user rating (fig.2), where the number of screenshots in iPad devices correlates the most. However, price shows almost no correlation with the user rating which is counterintuitive. In fact, the dataset contains more free apps (price = 0) than paid apps, causing the weak correlation between the user rating and the price (fig.3). As a result, apps were divided into free (price=0) and paid apps (price>0) for further analysis.

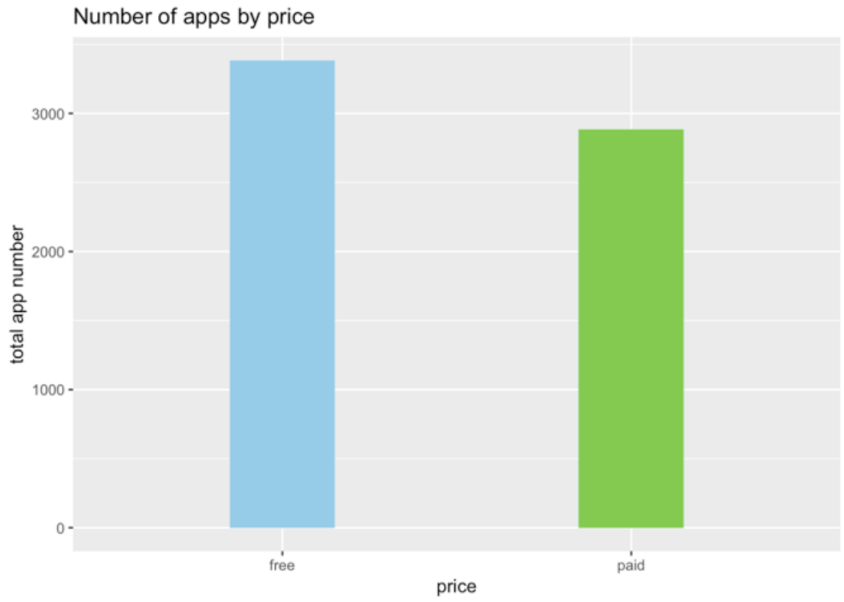

Figure 3: Comparison of app numbers for free vs paid

Resampling Data (Bootstrap)

In order to visualize how average user rating among different categories are affected by some correlated variables, we use the resampling method to estimate the mean values for the population's average user ratings from our sample and visualize it using the 95% confidence intervals that include the true population means.

Here, the user ratings among different categories were separated by correlated variables(price, number of languages, or number of screenshots) respectively. For the selected variables, which are correlated to the average user rating (fig.2), we separated those continuous/discrete variables into categories:

Price: Free vs Paid
Number of languages: Multi-language vs One language
Screenshots in Ipad Devices: 5 screenshots (maximum), some (1-4) screenshots, no screenshot

Bootstrap statistical inference was used in this study to gain the 95% confidence intervals of the dependent variable – average user rating (user_rating). The 95% confidence intervals were yield from 10000 resampling which goes from the 2.5th percentile to 97.5th percentile of the resample distribution.

Other than using the resampling methods, we could directly use the mean of the average user rating from our sample to estimate its values in the population for different genres, separated by either the price, the number of languages, or the number of screenshots. However, using the bootstrap resampling methods could help decrease the sampling error and make the estimation for the population parameters more accurate. Besides, we gained the 95% confidence interval using the bootstrap method to see whether or not the mean value of the sample average user rating is within the confidence interval. If the sample mean is within the 95% confidence interval, using the sample data for building the prediction model would give us a relatively accurate model for the average user rating in the population.

Prediction Model (AIC evaluation)
Before the feature selection, the distributions of all independent variables were examined to ensure the scale of variables are appropriate for the prediction model (fig.4). The distributions of some variables are highly skewed, therefore the values of these variables were transformed for better model prediction.

The feature selection was done by using AIC methods with the following variables. In our data analysis, the Akaike Information Criterion (AIC) is used to evaluate the relative quality of the model for the given dataset. The lower value of AIC is, the better the model is qualified for the given dataset.

Variables in the AIC methods:
y = user rating (0-5)
x_1 = number of screen shot (discrete)
x_2 = number of languages (discrete)
x_3 = size in MB (continuous)
x_4 = total rating count (discrete)
x_5 = number of supportive devices (discrete)
x_6 = price (continuous)

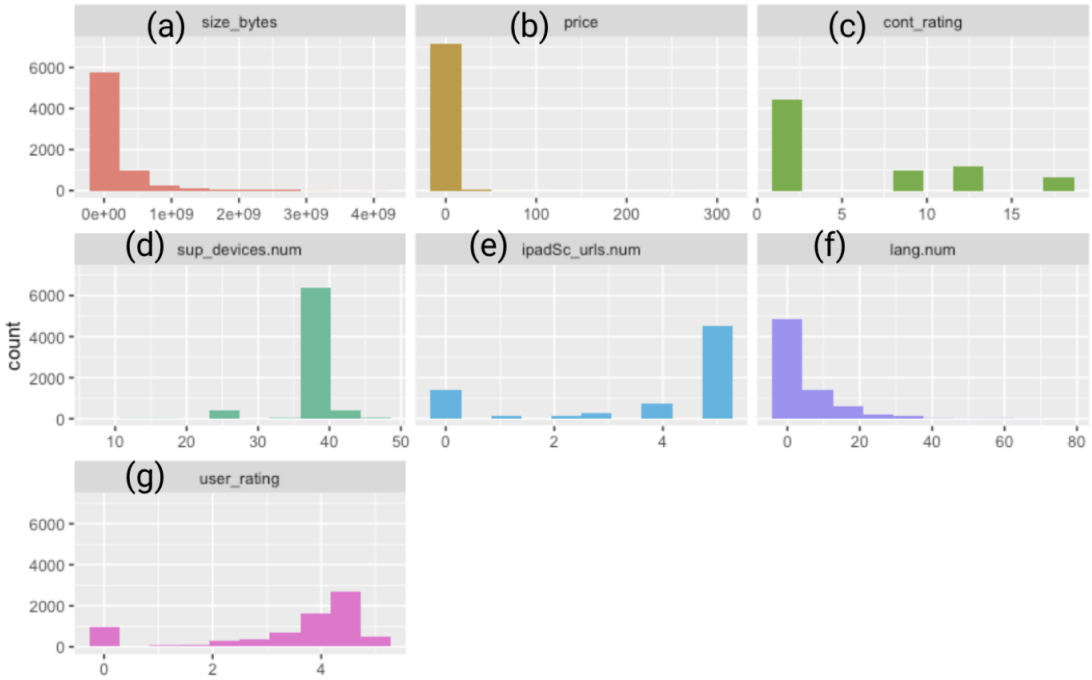

Figure 4: The distribution of the selected variables. (a) app size, (b) price, (c) content rating, (d) supporting device number, (e) ipad screenshot numbers, (f) language number, and (g) u ser rating.

In total, three models would be built for prediction: one for the whole dataset, one for free apps (x_6 was eliminated because all price is 0), and another one for paid apps (include x_6).

Below is the initial regression model for the whole dataset and paid apps using all variables (for AIC):

$$Y = w_0 + w_1 * transform(x_1) + w_2 * transform(x_2) + w_3 * transform(x_3) + w_4 * transform(x_4) + w_5 * transform(x_5) + w_6 * transform(x_6)$$

The initial regression model for free apps using all variables (for AIC):

$$Y = w_0 + w_1 * transform(x_1) + w_2 * transform(x_2) + w_3 * transform(x_3) + w_4 * transform(x_4) + w_5 * transform(x_5)$$

In the feature selection process, variables would be removed if eliminating them caused lower AIC values. This process was repeated until no more variables could be removed to yield a lower AIC value. Thus, three models would be built with selected variables. The accuracy of the prediction model would be

examined by R^2 value.

3 Result

Resampling Methods (Bootstrap)
The bar charts(Figure 5-7) show the 95% confidence intervals for the user ratings using black error bars gained from the bootstrap resampling and the mean value of the average user rating for each genre from the sample using color coded bars.

From Figure 5-7, we could tell that the mean values of the average user rating are gained from our sample are within the 95% confidence interval gained from the bootstrap resampling. As a consequence, the average user ratings were not obtained by chance and the dataset can provide sufficient information for the model prediction of the average user rating based on these selected features.

Prediction Model (AIC)
Here is the resulting model for the whole dataset with the lowest AIC score compared to other models with a corresponding R^2 score of 0.818, which means that 81.8% of the variance of the user rating could be explained by this model.

$$y = -0.3251 + 0.0003848x_1^4 + 0.0749x_2^{1/5} + 5.445 * 10^{-37}x_3^{10} + 3.481x_4^{1/50} + 0.00000282x_5^3$$

Besides the model for the whole data, here are the ones specifically for paid and free apps in the IOS mobile system. After separating the apps by their prices (price = 0 or price > 0), the AIC values for both models get lower, which means the model's quality gets better.

Free apps with R^2 of 0.8519

$$y = 0.7561 + 0.0002714x_1^4 + 0.1543x_2^{1/8} + 0.06206log(x_3) + 3.589x_4^{1/72} + 1.353 * 10^{-7}x_5^4$$

Paid apps with R^2 of 0.7457

$$y = -0.2885 + 0.002175x_1^3 + 0.05908x_2^{1/3} + 1.509 * 10^{-58}x_3^{16} + 3.557x_4^{1/49} + 0.03899x_6^{1/3}$$

For the model of free apps, since AIC did not change whether we put x_5(the number of supportive devices) or not, we excluded it, and for the paid apps, because the price varies, we included the price variable in the model.

With our models, we can now predict the user rating for any apple apps with variables we have for each model. For example, let's say you want to predict the

user rating for a paid game app ($0.99) with the 3 screenshots in ipad Itunes stores, 4 usable languages, size of 34.4 MB, total rating count of 6334, and 10 supportive devices. Since this app should be paid, you would use the model for paid apps, and after you plug in given values in the model, you would get 4.15561 for user rating prediction.

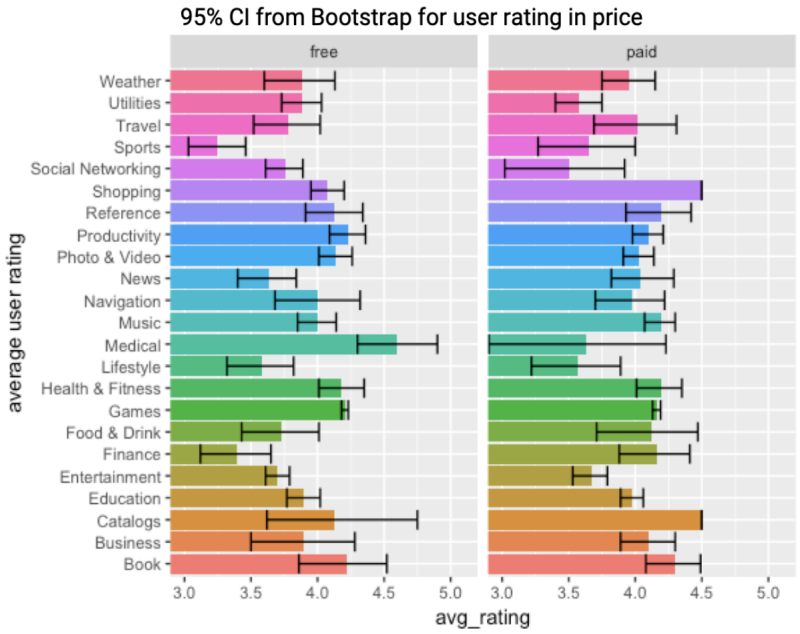

Figure 5: The relationships between the average user rating and the 95% confidence interval from bootstrapping. The dataset were separated by price (free/paid). Each of the bar charts represents the average ratings of the apps within different categories and are color coded based on the genre. The 95% confidence intervals were represented by the error bars.

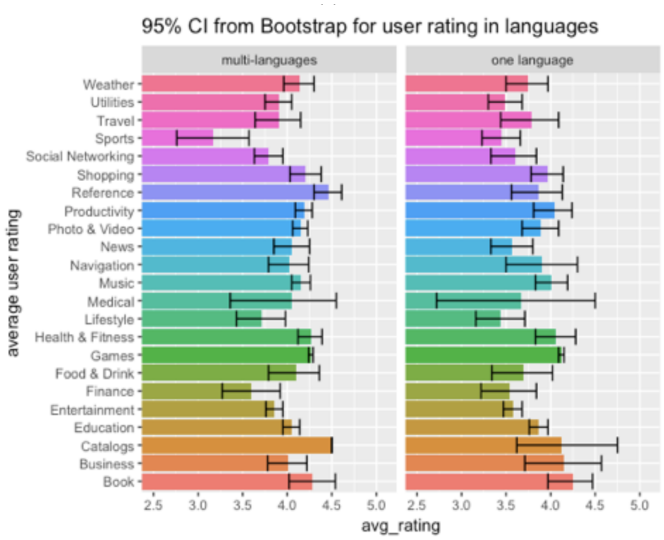

Figure 6: The relationships between the average user rating and the 95% confidence interval from bootstrapping. The dataset were separated by the number of language (one language/ multiple language).

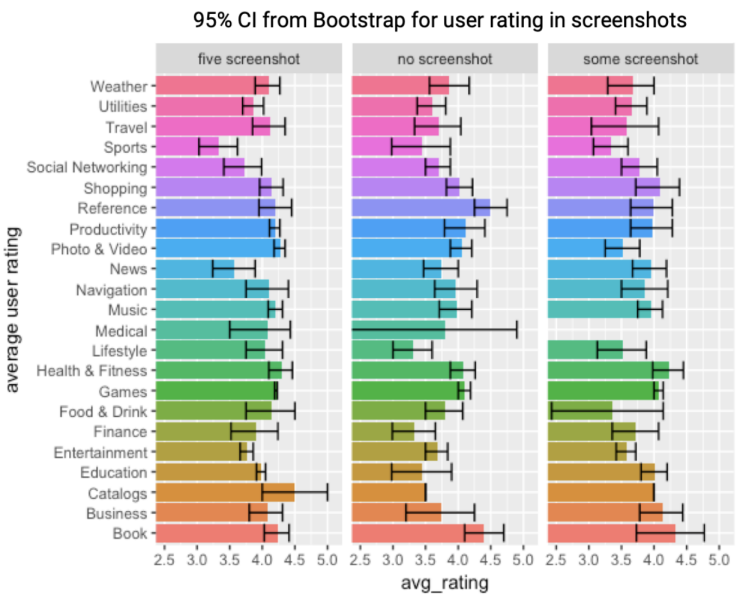

Figure 7: The relationships between the average user rating and the 95% confidence interval from bootstrapping. The dataset were separated by the number of screenshots (five screenshots/ no screenshots/ one-four screenshots).

4 Discussion

Even though the bootstrap method has provided a good estimation of confidence intervals of true means of user rating for each category, since the original data is not treated as the population, there could be errors regarding the true population mean of user rating. The reason for using bootstrap is that sometimes it works better than other methods with given data. As the data gets bigger, we believe that the possibility that the bootstrap method would perform precisely will get higher.

In addition, our main purpose is to predict the user rating of an app; however, not all people leave reviews or rate apps they've used. Even though we could make good models, in the real world, the prediction using our models might be inaccurate due to not having all people's opinions in the dataset.

5 Reference

1. Kozák, L. R., Dávid, S., Rudas, G., Vidnyánszky, Z., Leemans, A., Nagy, Z. (2013). Investigating the need of triggering the acquisition for infant diffusion MRI: a quantitative study including bootstrap statistics. *NeuroImage, 69,* 198–205. doi:10.1016/j.neuroimage.2012.11.063