Springboard Data Science Career Track
1st Capstone Project Report
Student: Huan Xia

Title: What makes a trending app on google play store?

1. Problem Statement:

Smartphone apps are changing our lives and are gaining more and more attention from people everyday. Based on data from statista.com, over 6000 new smartphone apps are produced everyday. However, most of them “died off” before they even get a chance to show up in our smartphone app search engine (e.g., Google Play store). So, how could a cellphone app survive such brutal competetion? Or what qualities must it possess to eventually be installed by customers? In this capstone project, real-world datasets collected from Google Play Store have been analyzed to provide insights to these questions. A Multiple or multivariate-regression model has also been developed in attempt to predict the rating of an app. The analysis outcome from this project may benefit software/app designers/engineers to adapt the app design in an effort to increase the likelihood of receiving a positive review.

2. Methods and Approaches:

2.1 Data source

The dataset can be accessed and downloaded from
https://www.kaggle.com/lava18/google-play-store-apps.
The contents of the datasets are scrapped from the google play store, which consists of app info of thousands of active apps. The first dataset includes basic app info, such as the application names, app categories, numbers of installs by customers, app ratings, app reviews, app sizes, app prices, etc. The second dataset includes the review sentiment information for different apps, such review sentiment polarity, number of reviews, review sentiment subjectivity, etc. Both datasets are well organized with different app/review parameters separated by columns. However, data cleaning is still needed to remove nulls, empty inputs, and to convert some parameters into desired format (e.g., string to number) prior to data analysis.

2.2. Data wrangling

2.2.1 Missing, null and duplicate values

There are missing and null values, duplicate rows, as well as undesired data format in both two original datasets. For the App dataset, the missing values exist in columns:

app rating, type, content rating, current ver and android ver. For the Review dataset, missing values exist in columns: Translated_Review, Sentiment, Sentiment_Polarity, and Sentiment Subjectivity.

Dropna function was applied to both datasets to remove rows with missing values in any column. Drop duplicate function was also applied to remove duplicate rows in both datasets. For the first dataset, the key column is the Rating column, and 9360 rows are retained out of 9367 original rows after cleaning. For the second dataset, the critical columns include Translated_Review, Sentiment, Sentiment Polarity and Sentiment Subjectivity; 37427 rows are retained out of 37432 rows after cleaning. There is very minimal data loss and impact on data integrity from this data cleaning process. And the revised datasets have been asserted to have no missing data, or bad data (e.g., negative rating in App dataset).

2.2.2 Merging datasets

There are two separate datasets for this project: one consists of Google Play store app info, and the other consists of the review info for individual apps. Data wrangling and exploratory data analysis were performed on each individual datasets. Further, the two datasets were joined for multi-regression analysis.

Prior to multi-regression analysis, the two datasets were merged on the shared column: App column. The App column is the identity column for the first dataset, but not for the second dataset (Review dataset), because one app can have many different reviews. And the Review column is the identity column for the second dataset. The two datasets were left joined on App column with the Review dataset being the left dataset.

The merged dataset was examined for nulls, missing values and duplicates prior to analysis.

2.2.3 Converting data type

The values in original datasets were primarily stored as string type, which can not be directly used for data analysis. Therefore, the data has been further cleaned and converted from string type to numeric or datetime type: e.g., removing dollar signs under price column, removing megabyte signs under size column, datetime parsing, converting app installs from string to integer, etc.

2.2.4 Data distribution, transformation

There are seven numerical columns in these two datasets after data cleaning and converting. The numeric columns are: App rating, App size, App reviews, App installs, App price, Review Sentiment_Polarity and Review Sentiment_Subjectivity. All of these seven data columns were checked for normal distribution to determine if any transformation is needed.

None of data in the numeric columns of App dataset were normally distributed. Log transformation was applied to App rating, App size, App reviews, App installs and App price. QQ plots were plotted and Shapiro normal tests were performed on the

transformed data. However, these transformed distributions are still not normal. In addition, exponential transformation was applied to App rating column, which converted the data to normal distribution.

The two numeric columns: sentiment polarity and sentiment subjectivity in the Review dataset are both normally distributed, and no data transformation is needed.

2.2.5 Outliers

Boxplots were produced for Sentiment_Polarity, Sentiment_Subjectivity and the exponentially transformed Rating data. According to the boxplots, there are outliers in transformed rating data and sentiment_polarity data. The outlier data points are defined as the data with a difference from the sample mean larger than twice the standard deviation. The outliers were identified and upon review of the ‘outliers’, it is concluded that these suspected outliers do not seem to be 'wrong' or false inputs. Without seeing the detailed review of the suspected rating outliers, we have no strong reason to exclude them. Also, the number of the suspected outliers was very small, which will not affect the data analysis if not excluded. Therefore, outliers identified in both datasets are kept for further data analysis.

3. Exploratory Data Analysis

Preliminary analysis was performed on both App dataset and Review dataset individually. And an exploratory analysis was then performed on the merged dataset.

3.1. App dataset

For App dataset, factors that may affect the app ratings and installs are explored.

3.1.1 App ratings

The mean of the rating generally fluctuates between 4.0 to 4.5. Some apps have really high ratings of >= 4.9, while other apps are rated below 2.0. According to the boxplots, there is not much difference in the mean of app rating among different categories. The Family category has the highest number of highly rated apps (>4.5), followed by Game category, then Medical and Tool category, respectively. This is mainly because that Family, Game and Tool categories also have the highest number of apps. If normalizing the highly rated app numbers to the total app numbers, we can see that categories with the highest percentage of highly rated apps are Event (0.55 %), Health and Fitness, then Parenting (0.5%).

When plotted against the number of installs, the average app ratings initially declines with the number of installs, for those with 1 to 10,000 installs. These apps with 1 to 10000 installs may be classified as “the unpopular apps”. And for “the popular apps”, those with above 10,000 to 100,000,000 times of installs, the average rating generally increases with the number of installs. And for those “most popular apps”, the one with

over 100,000,000 times of installs, the rating generally declines with the number of installs.

3.1.2 App installs

The most installed apps on Google Play stores are Subway Surfers (a game app) and Google News, both with over 1 billions downloads. The follow-ups are with over half a billion downloads and most of them are under Game category (40%) or Tech related (50%). All of these mega popular apps are targeted for audience of all ages, and all of these apps are free of charge.

The least installed apps are Ra Ga Ba and Mu. F.O., with only 1 download. Both of these two app are not free. The second least downloaded app with less than 5 downloads, are predominantly under the Game and Medical category.

3.1.3 App prices

The prices of the apps range from free to as much as $400. Over 92% of the apps are free, and nearly 90% of the highly rated apps (with rating >4.5) are free. The most pricey apps are under Lifestyle and Finance category, and the most expensive 20 apps only scored an average rating of 3.925, well below the average rating of all apps, with up to 50,000 times downloads.

3.1.4 Prices vs ratings

Linear regression analysis was performed on app prices and app ratings, and a weak and negative correlation between price and rating was observed (slope = -0.001 and p value=0.009). A weak correlation was also observed between log transformed price and app rating, with not much improved R value.

3.1.5 Prices vs installs

There is no significant correlation between app price and installs, but there is a weak, negative but significant correlation between log transformed price and installs, therefore making the apps free of charge or cheaper can potentially attract more installs.

3.1.6 Reviews and app size vs ratings and installs

Correlation matrix analysis was performed on app review, size, rating and installs. And individual regression analysis was performed on each pair of the four variables. According to the analysis, there is a very weak and positive correlation between the app review times and rating. Highly rated apps tend to have more reviews.

There is also a very weak positive correlation between the app size and rating. It is likely that the bigger the app, the more comprehensive and well-designed an app may have to win higher ratings.

Also, app size, numbers of reviews and installs are positively correlated, meaning bigger app leads to more installs and more reviews.

The preliminary approach is to 1) explore within each category, and examine for potential correlations among these different parameters, more specifically, their impact on users downloads and ratings, and identify the key factors to a popular app; 2) extract a subset of data prior to data analysis to be used for prediction validation purpose; 3) use previously identified key parameters to predict the user downloads and ratings of an app and validate against the training data.

3.2. Review dataset

For Review dataset, factors that may affect the sentiment polarity are explored.I Sentiment polarity describes how strong a positive or negative the review is. And the Sentiment subjectivity describes how subjective the review opinion is. The lower subjectivity the review has, the more unbiased the review tends to be. So both sentiment polarity and subjectivity can potentially be correlated to the actual app rating.

3.2.1 Sentiment polarity and subjectivity distribution

According to the density plot, most of the reviews tend to be positive than negative. And the reviews dominantly fall between slightly negative to highly positive. There is also a spike at polarity of 0.0. This could be due to the shortcoming of the natural language process: people use drastically inconsistent and versatile words and phrases to express their emotions and opinions in reviews, it is very likely that the methods used during the natural language processing is insufficient to correctly identify the true polarity of the review. Therefore, many reviews may not be properly processed and labeled, resulting in a high abundance of 0 polarity.

The majority of the subjectivity scores falls between 0.2 and 0.8 (very subjective to very objective). And there are also two spikes (extremely subjective and extremely subjective) observed at 0.0 and 1.0, which could also be contributed by the vocabulary insufficiency issue during the subjectivity natural language processing.

3.2.2 Sentiment polarity and subjectivity vs. category

Game, News and magazine, and Social apps have the smallest polarity ranges (-0.1 to 0.1) and most outliers, and their reviews tend to be neutral or mildly positive. Weather, personalization, books and beauty apps have the widest polarity ranges (spanning a range of up to 0.7), with the least outliers and the reviews being mainly mildly positive to positive.

For most categories, the reviews are moderately subjective, and lie between 0.3 to 0.7. Two categories, Beauty and Comics, have the widest subjectivity range, from extremely objective (0.0) to fairly subjective (0.8). The apps under these two categories are more likely to receive objective reviews than the rest.

3.2.3 Sentiment polarity and subjectivity vs. app rating

The merged dataset was used to perform the regression analysis between review sentiment polarity and app rating. According to the analysis, there is a significant positive correlation between review polarity and app ratings: the higher review polarity, the higher ratings.

Interestingly, review subjectivity is also positively correlated with app ratings: the higher subjectivity, the higher ratings.

3.2.4 Predicting rating using review polarity and subjectivity

As both sentiment polarity and subjectivity are positively correlated with app rating, a multiple linear regression analysis was performed using review polarity and subjectivity to predict app rating. The final equation: Rating = 3.889 + 0.4* Polarity + 0.61 * Subjectivity, only obtained a R2 score of 0.08, with 1 being the perfect prediction. Therefore, the multiple linear regression equation did not provide a very accurate prediction of App rating. In order to improve the prediction, a non-linear regression or data transformation of the variables may be investigated in the next steps.

4. Hypothesis Testing

According to our previous correlation analysis, several trends have been noticed:

1) free apps tend to have more download;
2) bigger size apps tend to have higher ratings;
3) app category may affect the review sentiment and eventually ratings.

The price, size and the genre of an app are all factors that can be controlled by app developer and design companies. Although weak correlations were observed in previous correlation analysis, several hypotheses have been tested to confirm the above observations:

4.1 Do free apps tend to have more downloads?

A two sample bootstrap test was performed to examine this hypothesis. The p-value of 0.0 suggests that free apps do have much more downloads (up to 4,758,709 times more) than paid apps.

4.2 Do bigger size apps tend to have higher ratings?

A two sample bootstrap test was also performed to examine this hypothesis. The cutoff between big and small app is the median app size of 13 mb. And the p-value of 0.0 suggests that apps with bigger size (>13 mb), and potentially more comprehensive with more features do receive higher ratings (of 0.043) than small apps (<13 mb) .

4.3 Does app category affect the review sentiment?

According to our previous analysis on Review dataset, apps under different categories may have different review sentiment polarities. When designing an app, one can classify the app under alternate categories. Different category may have different target audience with different sentiment characteristics. In this case, does which category the app is placed under affect the review? For example, a weather forecasting app can be placed under either Weather category or Tool category. Here, we tested that weather the review sentiment between the two categories is different using two sample bootstrap test. The p-value of <0.01 suggests that there is a difference in app review sentiment among different app categories, with weather category having 0.155 higher positive sentiment reviews than tool category. Therefore, placing the weather forecasting app under weather category may benefit the app rating than placing it under the tool category.

5. Recommendations:

To gain a higher app rating or downloads, the app designer may consider making the app free of charge, making the size bigger and equipped with more features and functions. And if possible, the app designers should consider placing the app under which category, or designing the app for an audience with a generally more positive review sentiment.