Google Play Store App Analysis

Final Report

Najja Osiomwan

November 16th, 2020

Google Play

Table of Contents

Executive Summary

The purpose of this project is to perform an exploratory data analysis (EDA) of the Google Play Store dataset. The insights derived from this analysis can be used by app developers to learn what factors are most important in creating a successful app. The dataset consists of information on over 267,000 apps. I cleaned the data and formatted it for analysis. Data analysis and statistical tests were performed to identify trends and understand how consumers interact with apps in terms of ratings, reviews and installs. *InstallCat*, a binary variable created to categorize the number of installs an app received as high or low, was chosen as the target variable for logistic regression. Using stepwise regression, I was able to find the best multiple regression model to use for predicting the *InstallCat* outcome. While the results here are significant, I do believe that more data is necessary in order to make a more accurate prediction. The results of my findings are presented in this report.

Problem Statement

Developers spend months of time and energy trying to bring some of their greatest ideas to life in the form of a mobile app. With the increase in use of mobile devices there has been a huge explosion in the number of apps available for entertainment, productivity and to ease the life of consumers in general. As of August 2020, approximately 111,000 apps are released on the Google Play Store each month[1]. A Gartner study also reveals that the success rate for a commercial app in the market is just 1 in 10,000[2]. With so much competition in the market, app developers can benefit from some form of analysis to help determine some of the key features that contribute to an app's success. The analysis in this report can help app developers create an app that will gain high ratings and installs. It can also be used by investors to help understand the likelihood of success for an app they plan to invest in.

Problem Solving Approach

In this report I will take a deep dive into the data available on apps from the Google Play Store. I closely look at each variable associated with these apps, check for normal distribution and learn common features and anomalies on the apps store. I form a few hypotheses about the data and test their significance to develop insights about how users interact with these apps. Finally, I create a logistic regression model that is optimized to give the best prediction on whether an app will receive a high or low number of installs based on the features available in this dataset.

Data to Be Used

The "**Google Playstore Full.csv**" data can be found for free on Github at the link in the appendix[3]. The link contains data on over 267,000 apps. In attempt to analyze the dataset and find a model to predict the installs an app will receive, the following variables were used:

1. Installs (target variable): The number of installs an app received.

2. Rating: The rating of an app on a scale of 1-5.
3. Reviews: The number of reviews an app has received.
4. Category: The category of app based on subject matter. (Educational, Gaming, etc.)
5. Size: The size of an app in bytes.
6. Price: The amount of money charged to download the app.
7. Content Rating: Rating of the content on the app. (Teen, Mature, Everyone, etc.)

During the data cleaning process, there were some variables that I decided not to use for analysis. These variables include *App Name, Last Updated, Minimum Version*, and *Latest Version*. These variables do not provide enough information to use for analysis. There were also new variables created in order to change numeric variables into categorical variables. These variables were *Size Groups*, which grouped size into five categories, and *InstallCat*, which categorized the number of installs into high and low.

Data Cleaning

The data wrangling approach and techniques used to prep the data is described in the following steps.

a. Imported the dataset and necessary libraries for analysis.
b. Reviewed a summary of the data and removed 16 rows and 4 columns that had missing data or misaligned data.
c. Removed symbols from variables that included plus signs, dollar signs, phrases such as “and up”, and numeric abbreviations such as k or M for thousand and million respectively.
d. Converted *Last Updated* variable to date format although I realized later, I would not be using this variable.

Univariate Analysis

It was important for me to understand the distribution of each variable I planned to use during the EDA before I started testing for significance. In analyzing each variable individually, I realized some of the data needed to be either excluded or transformed in order to achieve statistical significance in later sections. The following are the results of my univariate analysis. A more detailed report can be found in the Exploratory Data Analysis Report[4].

Rating: This continuous variable was converted to a numeric data type in R. The data was heavily left skewed, so I decided to remove outliers from the dataset (~3% of the data). After removing outliers, the bell curve of the Rating variable more closely resembled a normal distribution.

Reviews: This integer variable was converted to a numeric data type in R. The data was right skewed as a few apps receive a large number of reviews. By using a log transformation of the data, I was able to find a normal distribution.

Category: This categorical variable was converted to a factor data type in R. The variable originally had 68 levels. I grouped these into four categories in order to make the data more manageable and proportional. The four new categories and their proportions are:

a. Education (29%)
b. Entertainment (25%)
c. Lifestyle (22%)
d. Productivity (24%)

Installs: I decided that, although this data should have been numeric, it would be best to convert it to a categorical data type. By categorizing *Installs* into five groups I was able to resolve the heavy left skew. The groups and their proportions are:

a. 0-500 (15%)
b. 1000-10000 (28%)
c. 10000-100000 (32%)
d. 1000000-1000000 (18%)
e. 1000000+ (6%)

Size: "Size" contains mostly numeric values as well as a "varies by device" value. Because of this, the variable had to be categorical. I grouped the values into the five buckets below:

a. Varies with Device (4%)
b. 0-4M (24%)
c. 4M-8.3M (24%)
d. 8.3M-19M (24%)
e. 19M-334M (23%)

Price: This variable was disproportionate as only a 5% of the apps have a fee to download. This caused me to convert this numeric variable to a categorical variable with two groups:

a. Free (95%)
b. Paid (5%)

Content Rating: This categorical variable originally had 5 groups. Because one group, "Everyone", consisted of 90% of the data, I consolidated all other groups into one group called "Mature". The proportions are below:

a. Everyone (90%)
b. Mature (9%)

Bivariate Analysis

Because *Installs* is the target variable for this project, I looked into how each variable effected the amount of installs an app received. I ran a series of statistical tests on the variables *Size*, *Price*, *Content Rating*, and *Category* and found that they did not have significant effects on *Installs*. The tests that were statistically significant are presented below:

e. Ratings & Reviews: These variables have a weak correlation of -0.17. This tells me that I can include both ratings and reviews in my regression model since multicollinearity is not an issue here.

f. Installs & Rating: Analysis of these two variables show that apps with fewer installs have higher ratings.

g. Installs & Reviews: Installs and Reviews have a positive correlation. As the number of installs increases the number of reviews increases as well. The box plot presented in the detailed report shows that apps with over 1 million installs have the highest number of reviews on average.

A more detailed report of these findings can be found in the Statistical Analysis report[5].

Hypothesis Testing

By performing hypothesis tests, I was able to learn more about how consumers interact with apps on the Google Play store. A more detailed report can be found in the Statistical Analysis report[5]. A summary of my findings is presented below:

h. Hypothesis: On average, education apps will have higher ratings that apps in other categories.

- i. The results of the ANOVA test show that we can reject the null hypothesis. Education apps are rated higher than all other categories of apps. Because education apps help people improve their skills, it makes sense that consumers who have a positive experience with an app and are actually learning from the app are more likely to take the time to leave a 5-star rating

i. Hypothesis: On average, Education apps will have a higher number of reviews than Lifestyle apps.

- i. The p-value for this one-sided t-test is 1. I accept the null hypothesis at the 95% confidence level and determine that there is no difference in number of reviews for these two categories of apps.

j. Hypothesis: On average, Paid apps will be rated higher than Free apps.

- i. The p-value for the ANOVA test is less than 0.05 so I can reject the null hypothesis at the 95% confidence level. Paid apps have a slightly higher rating than free apps on average. This may be because consumers who pay to download an app would be more satisfied with the app and rate the app higher.

k. Hypothesis: Apps Rated Mature will have more reviews than apps for Everyone

i. The p-value for the chi-squared test was less that 0.05 so I reject the null hypothesis and conclude that apps rated mature have a higher number of reviews on average.

Regression

For regression, *InstallCat* was used as the target variable. This means that the goal of the model is to predict if an app would receive high installs (greater than 100k) or low installs (less than 100k). Because the target variable is categorical, I used logistic regression. Forward selection, backward elimination, and stepwise regression methods were used for variable selection.

The final model included all relevant variables and all variables were statistically significant at a 95% confidence level. The McFadden's R-squared value was 0.58 which tells me that the model explains a little more than half of the variance in the *InstallCat* variable. A more detailed analysis can be found in the Statistical Analysis report[5].

Conclusion

In sum, the data provided in the Google Play store dataset gives little insight into how users interact with apps on the apps store. We learn that paid apps as well as apps in the education category are more likely to receive higher ratings and in the education category and apps rated mature receive more reviews that other apps on average. We also learn that *Ratings* has a negative relationship with number of installs while *Reviews* has a positive relationship with number of installs. However, when it comes to predicting how many installs an app would receive, we need more data. With the current model we can only explain about 58% of the variance in *InstallCat* – the install category. It is also important to note that logistic regression was used on the binary variable *InstallCat* because the *Installs* variable was extremely right skewed. Many of the variables in this dataset were either skewed or disproportionate. Having data with a more normal distribution could potentially increase the number significant test results in this report. An added benefit is that it would allow me to use a linear regression model that may be a better fit for prediction of installs as a numeric variable. Including variables that explain how users interact with the app after the install may also be beneficial for this analysis. While calculating the accuracy of the model is outside the scope of this project, the regression model and the report could be improved by testing the model on sample test data.

Appendix

1. https://www.statista.com/statistics/1020956/android-app-releases-worldwide/
2. https://www.businessofapps.com/insights/top-reasons-why-mobile-apps-fail-to-make-a-mark-in-the-market/
3. https://github.com/NajjaOsiomwan/GPS
4. GPS Exploratory Data Analysis.pdf
5. GPS Statistical Analysis.pdf