A Scalable Approach to Predict the Profitable App Categories from the Google Play App Store Data

[Extended Abstract]

Rayhan Hossain
EECS
University of Tennessee, Knoxville
rhossai2@vols.utk.edu

ABSTRACT

People are becoming more dependent on mobile applications gradually, and the mobile app market is growing rapidly. To help the startups whose primary goal is to develop Android applications and generate revenue from them, we build a model which can predict the profitable app categories analyzing the previous data of Google Play store. A scalable approach using the MapReduce [2] with the support of pyspark is applied for processing and computing the result. Finally, it illustrates the significant findings with meaningful charts and provides the guideline for analyzing the local app market.

Keywords

MapReduce; KMeans; Correlation

1. INTRODUCTION

The smartphone is one of the most significant innovations of the century which has developed our lifestyles in numerous ways. Not only for managing our principal office and personal jobs but also to entertain ourselves it plays a meaningful role. As a result, the app market is also overgrowing. Being an open source platform and easily accessible to developers, the Android app market has risen around 217 percent in last two years. Considering the app market size, more developers are shifting interest in Android development, and the number of Android app development based startups are increasing. When a startup comes into the market, it requires to investigate the app market and determine the future business direction. Developing what kind of applications might be beneficial for business is the obvious question for each startup. To assist the startups we tried to answer some critical questions including the following four- Which App categories have the maximum app downloads? Which one might be profitable for business, fee apps or paid apps? Which age group of users should be the target of the business? How many different Android versions should be supported to maximize the number of app installs? To define the term "profitable" we consider only the number of total app installs. As we cannot get insight into the revenue generated from the Ads, the number of total installs is the only prime factor which dominates the income. We explore the Google Play app store data in a scalable way and discover the relationship between the number of total installs and other factors. Finally, we present the outcome using mixed charts which benefits the startups to understand the Android app market.

Table 1: Data Column and Data Type

Data Column	Type
App	String
Category	Categorical
Rating	Numerical
Reviews	Numerical
App Size	Numerical
Total Installs	Numerical
Type	Categorical
Price	Numerical
Content Rating	Categorical
Genres	Categorical
Last Updated	Date
Current Version	Categorical
Android Version	Categorical

2. METHODOLOGY

We collect our dataset from kaggle.com [1]. There are 10,814 data points with thirteen attributes for each row. Five of the attributes are numerical values, and six of them are categorical (Table 1). Another two is not relevant for our analysis.

Before starting the analysis part, it is required to clean the data. We found some unexpected characters for the numeric columns in the chosen dataset like-comma, +, extra spaces, etc. which are resolved in the cleaning part. This cleaning task is handled by writing a Python function. After that, the finalized data are

sent through the MapReduce process in the format of Spark RDDs. Once the key-value pairs are generated using the Map function, we apply the Reduce function for different key columns. Mostly we use the relevant categorical columns as our key values like- App Categories, Type, Content Rating, etc. Finally, we display our results with mixed of charts using matplotlib. Our workflow is shown briefly in figure 1.

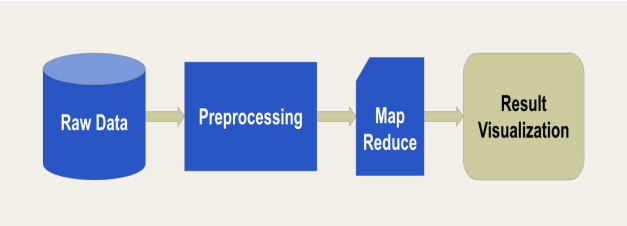

Figure 1: Work flow

Additionally, we also generate a heat map of the correlations among numerical data columns to understand the data dependency. The heat map indicates that app review has a high association with the number of installs. However, we didn't find any significant dependence among other numerical attributes. Considering this, we determined to analyze the app installs with respect to the categorical data.

3. EVALUATION

Our test data set is pretty small, and the processing time using pyspark was negligible. Mostly the number of app installs is related with the categorical data columns like- App Categories, Type, Content Rating, and Android Versions. In the beginning, we compare the number of total app installs with app categories. Our hypothesis was, game and social networking applications should dominate the app market. The result coincided with our interpretation. The most installed app categories were- Game, Communication, Social, Photography, and Tools.

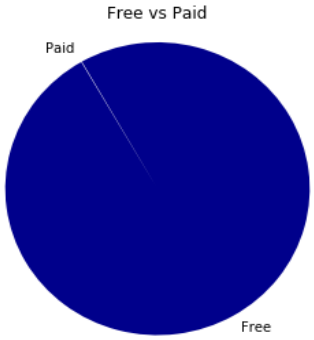

Figure 2: Number of total installs- Free vs Paid

Later, we desired to answer the second question regarding free vs. paid apps. Here we discovered that the total number of installed free apps was 167560477406 whereas the number for paid apps was 72956081. So, only 0.0435 percent of the total installed Android apps are paid (Figure 2). From this, we can conclude that starting a startup with paid application might not be a great idea.

Targetting the users from all ages is always a safe idea. However, if any startup wants to develop apps for a specific user group than picking the teenage would be advantageous for business (Figure 3).

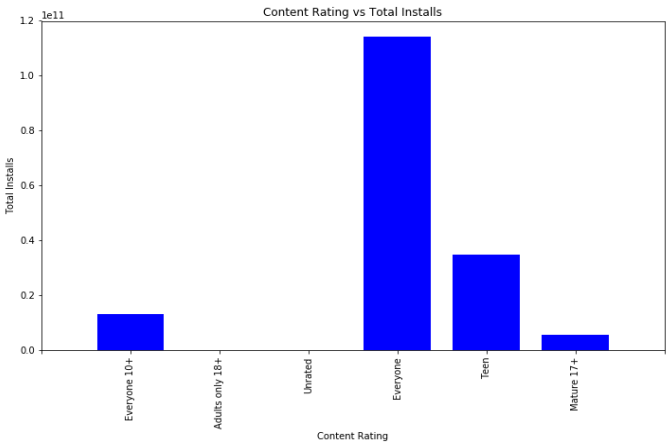

Figure 3: Content Rating vs. Total Installs

Finally, it is not necessarily valid that you need to support all the available Android versions to maximize your downloads. The result reveals that if the developer supports the devices with Android OS version 4.0 and up, they will meet above 97

4. CONCLUSION AND FUTURE WORKS

Using our small dataset, the results we found are pretty convincing as they resemble with most of the hypothesis. Since the model is scalable, we can employ it to analyze the massive dataset of the Google Play app store. Applying our design, we are purposing to build a web application which will assist the startups to investigate their local market as well as the global one.

5. REFERENCES

[1] https://www.kaggle.com/lava18/google-play-store-apps

[2] J. Dean and S. Ghemawat, "MapReduce: simplified data processing on large clusters, "Communications of the ACM", vol. 51, no. 1, pp. 107-113, 2008.

[3] M. Bodoia, "Map-Reduce Algorithms for k-means Clustering."