Analysis of Google Play Store Data set and Predict chances of an app being successfull on Google Play Store

Rimsha Maredia
Texas A&M University
College Station, Texas
rimsha.maredia@tamu.edu

Abstract

The google Play Store is the largest and most popular Android app store[2].We have found a raw data set of Google Play Store on kaggle App. It has enormous amount of data that has potential to make an optimal model.It contains 13 different features that can be used for predicting how successful a new app will be using these features[1].This data set is scraped from the Google Play Store.

We propose a model that would predict chances of any app getting successful by predicting different features like app rating,number of downloads etc. We will be using Google app Store data set and analyse the success of all apps from different categories from Google Play Store[1]. Our aim is to evaluate this data set and make a model that would drive app making business to success[4].

1. Introduction

The purpose of our project is to gather and analyze detailed information on apps in the Google Play Store and get an insight on app feature and the current state of the Android app market. In our data set we are focusing on features like Category, Rating, Installs,Current Version,Category,Price and Genres[1].In order to make an optimal model,We will also be using most relevant app reviews, sentiments and sentiment polarity score for analysis[5].There are over 30000 apps across 11 different categories in the Google App Store. Each app has its own web page where detailed information is available[4].

Most of the apps in the Google Play store are found free while installation. Some of them have in-app purchases feature that refers to buying extra content inside an application on a mobile device.The overall flow of analysis consists of following steps: data extraction from kaggle data set,data cleaning, correlation and cluster analysis,splitting training and testing data set and using different machine learning models to make an efficient model for prediction.

Our key observation at first glance includes how we can use this data set to increase overall demand of Google Play Store and get to know what kinds of apps are in more demand and which apps are not so that we can decrease the time spent on apps which are not popular on Google Play Store. If we look from a company's perspective that is trying to launch new app, we need to look at the past failures to avoid mistakes[6]. Our goal is to decrease these mistakes by analysing both successfully and unsuccessful apps. In our data set, we have 'most relevant' 100 reviews of each app and they are pre processed with following 3 new features - Sentiment, Sentiment Popularity and Sentiment Subjectivity. This data can be used to improve the performance and get more business values[1].

After looking at various data sets on Kaggle web page. We have found a data sets which would fit according to our requirements and the features in the set would result into an optimal model.Our experimental analysis will validate features that effects google app business and if we are successful and are able to make an optimal model, it can be used for future work[6].

2. Preliminary Literature Survey

There have been numerous application of machine learning in the industry;Amazon store, IBM e-commerce and others have employed machine learning in product classification as well as product recommendation.Numerous research has been done to improve classification in various domain; an example is a research done by Schnack et al. [7] using machine learning to classify patients with schizophrenia, bipolar disorder and healthy subjects with their structural MRI scans.

The statistics shows that the number of applications in the Google Play Store have increased exponentially from December 2009 to December 2019 [2]. The need for creating applications that effectively targets and grows a user's base has been a growing problem in recent years. As the

novelty of phone applications wear off, it becomes increasingly difficult to garner support for new apps. As such, the need for more research involved with determining the chance of success in an application is increasingly important. Almost all aspects of business in modern times require a successful smartphone application[3].

3. Proposed Technical Plan

3.1. Data Set

We are planning to use Google Play Store Data set for both training and testing our model. We will split it into two parts, 70 % (training) and 30 % (testing) the model[5].In addition, we have data set of 100 most relevant reviews which are attributed with 3 features-Sentiment, Sentiment Polarity and Sentiment Subjectivity[5]. We will also be using them for training purpose.

3.2. Analysing different features

The factors that require attention in solving this problem are: (1) Category: Category that app belongs to like Beauty, Business etc. (2) Rating:Overall user rating from 1-10. (3) Reviews:Number of user reviews. (4) Size:Size of each app, how much memory they consume. (5) Installs:Number of installs for the app. (6) Type: App is paid or free. (7) Price:Price of each app, 0 for apps that are free. (8) Content Rating:Age group the app is targeting, will help in finding the relation between age group and number of apps in demand. (9) Genres: Shows what genre a particular app belongs to. For example, Business,Music etc. An app can belong to more than one genre. (10) Sentiment:Positive/negative/neutral. Will use it only for training. (11) Sentiment Polarity:Polarity score, will be using only for training. (12) Sentiment Subjectivity:Sentiment subjectivity score. Will be using only for training.

3.3. Data Processing and modelling

It is very important that each and every piece of raw data leads to a more accurate result. We will be first cleaning the data by removing duplicates,removing null values by mean values in the ratings column[1] and then converting them into appropriate forms etc.

As our data set contains many unique features and we have large number of features,therefore we will be using "Random Forest Regression" for training the model[4].We will also train the model using other models like Decision Tree and K-Mean clustering and see which model gives highest rate of accuracy on testing set.If an effective model is made, we will be using regression to predict the rating of the app.After this our aim is also to determine the kinds of apps that are likely to attract more users.

3.4. App Popularity and Staleness

In addition we will also be considering features like update and number of downloads and latest version to find its popularity.An important property of a market is its "activity", or how frequently are apps being maintained. We say that an app is stale if it has not been updated within the last year from the observation period, and active otherwise. We propose to use the download count to determine app popularity[6]. Analysing above features is what we will be doing as a part of this project.

3.5. Evaluation

This goal of our project is to create a model that can be used to get different statistics of the app market. This can help app business in several ways. If the app holds a feature that might change the future usage of users, a data-driven business venture could launch the app in the market to get a better hold of the market relying on the key features and marketing future development. We also want to analyse transition of app market in different time period.

References

[1] Kaggle.com.(2018). Google Play Store Apps.[online] https://www.kaggle.com/lava18/google-play-store-apps [Accessed 3 Mar. 2020].

[2] Google play store: number of apps 2018(2018). [online] https://www.statista.com/statistics/266210/number-of-available-applications-in-the-google-play-store/ [Accessed 3 Mar. 2020].

[3] Amit Chile, Dr. P. R. Gundalwar.(2019). Analysis of Google Play Store Application.[online] http://ijraset.com/fileserve.php?FID=24134 [Accessed 3 Mar. 2020]

[4] Harman, M., Jia, Y., and Zhang, Y. (2012). App store mining and analysis: Msr for app stores. In 2012 9th IEEE Working Conference on Mining Software Repositories (MSR),pages 108–111.

[5] Jong, J. (2011). Predicting rating with sentiment analysis. [online] http://cs229.stanford. edu/proj2011/Jong-PredictingRatingwithSentimentAnalysis.pdf.

[6] [2015].Grover, S. 3 apps that failed (and what they teach us about app marketing). [online] https://blog.placeit.net/apps-fail-teach-us-app-marketing/.

[7] H. G. Schnack, M. Nieuwenhuis, N. E. van Haren, L. Abramovic, T. W. Scheewe, R. M. Brouwer, H. E. Hulshoff Pol, and R. S. Kahn, "Can structural MRI aid in clinical classification? A machine learning study in

two independent samples of patients with schizophrenia, bipolar disorder and healthy subjects," NeuroImage, vol. 84, pp. 299–306, jan 2014.