JOURNAL OF DATA SCIENCE | Vol.2022:01
eISSN:2805-5160

Data Analysis and Rating Prediction on Google Play Store Using Data-Mining Techniques

Kayalvily Tabianan [1], Denis Arputharaj [2], Mohd Norshahriel Bin Abd Rani[3], Sarasvathi Nagalingham[4]

Faculty of Information Technology, INTI International University, 71800 Nilai, Negeri Sembilan, Malaysia

Email: kayalvily.tabianan@newinti.edu.my

Abstract

Google Play Store was formerly known as Android Market. This biggest Android Application (App) provides a wide variety of details on requirements such as reviews, quality, number of installs, and explanations for device functionality. This study aims to predict the ratings of Google Play Store apps using decision trees for classification in machine learning algorithms. The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by learning simple decision rules inferred from prior data. This method classifies a population into branch-like segments that construct an inverted tree with a root node, internal nodes, and leaf nodes. The algorithm is non-parametric and can efficiently deal with large, complicated datasets without imposing a complicated parametric structure. This enables us to draw a comprehensive picture of the current situation on the process of analyzing Google Play Store by Number of Downloading Rate and Rating in current market trend. This will help the developers understand customers' great desires, attitudes, and trends in demand. To understand more in-depth, the similarity between the functionality of the device and to construct clusters of related applications. Then, analyze their characteristics following features of interest. The datasets that the author used are collected from Google Play Store (2019). In this research, the expected results have a more strong correlation between price and number of downloads and similarity between price and participation.

Keywords

Google Play Store, Decision Tree, Analysis, Machine Learning Algorithm

1. Introduction

Google Play Store is the biggest platform where application developers showcase their applications. The google play store is one of the largest and most popular Android app stores. It has an enormous amount of data that can be used to make an optimal model. The number of applications on Google Play Store is rising drastically on every single day. However, not every application has become popular among Android users. Every application development undergoes systematic processes and these procedures consume time and funds. Therefore, by knowing, the number of downloading rates and review rates can help developers and business managers enable to obtain insights about profitable and non-profitable apps.

Submission: 8 November 2021; **Acceptance:** 5 January 2022

Developers and users play key roles in determining the impact that market interactions have on future technology. However, the lack of a clear understanding of the inner working and dynamic of popular app markets impacts both the developers and users. This will highly benefit the application developers in terms of reduce the chances of failure and able to achieve target goals. The proposed prediction system will be a great tool for the application developers because it consists various understandable charts developed from insights gathered by analyzing Google Play Store by number of downloading rates, rating and developed decision trees to predict the upcoming demanding applications.

2. Methodology

KDD Methodology

Knowledge discovery is the process of obtaining knowledge from various information and according to different needs. The purpose of knowledge discovery is to shield the user from the cumbersome details of the original data. It contains many different methods of discovery, including inductive learning, KDD Methodologies used in the proposed system. KDD is known as Knowledge Discovery in Databases. It has the highest number of steps compared to other methodologies, KDD is excellent for extracting the essence of information available, and that can generate reports, views, or summary of data for better decision-making, which has a large number of volumes of information. KDD is an iterative process where evaluation measures can be enhanced, mining can be refined, new data can be integrated and transformed to get different and more appropriate results. The cycle of information acquisition is repeated, collaborative, and made up of nine phases.

The first phase is to understand the production and awareness of the application client and sets the scene to consider what to do with conversion, algorithms, and representation. The second phase is to construct a data set, in this stage, the author has validated the data by determining whether it is needed to be clean or contained NULL value and choose among from all the data, which will help through the proposed system. The third phase is pre-processing and maintenance, in this phase, the author starts to clean dirty data such as cleaning the null value and eliminating the outlier’s value. The fourth phase is data transformation. This phase, involves dimension reduction, such as choice and elimination of features and record testing, and conversion of attributes such as discretization of numerical attributes and functional transformation. The fifth phase is choosing a suitable data-mining task, which is mostly, relies on the objectives. In this phase, whereby generalizing from enough training instances, a prototype is built explicitly or implicitly. The inductive approach's underlying assumption is that the learned template extends to future cases. The sixth phase is the data mining algorithm with the strategy, which is discovered in the previous step and decided with the tactics. This stage involves selecting the specific pattern search method, including multiple inducers. Lastly, this phase determines the research outcome and the enhancement of the systems.

3. Results and Discussion

Acceptance Testing

A questionnaire is a well-established tool within social science research for acquiring information on participant social characteristics, present, and past behavior, standards of behavior or attitudes, and their beliefs and reasons for action concerning the topic under investigation (Bulmer, 2004). The author has selected Google Forms as a platform to prepare the questionnaires. Google Forms can be time-efficient when the questionnaire is distributed. The respondents are application developers and a total of 35 respondents have participated

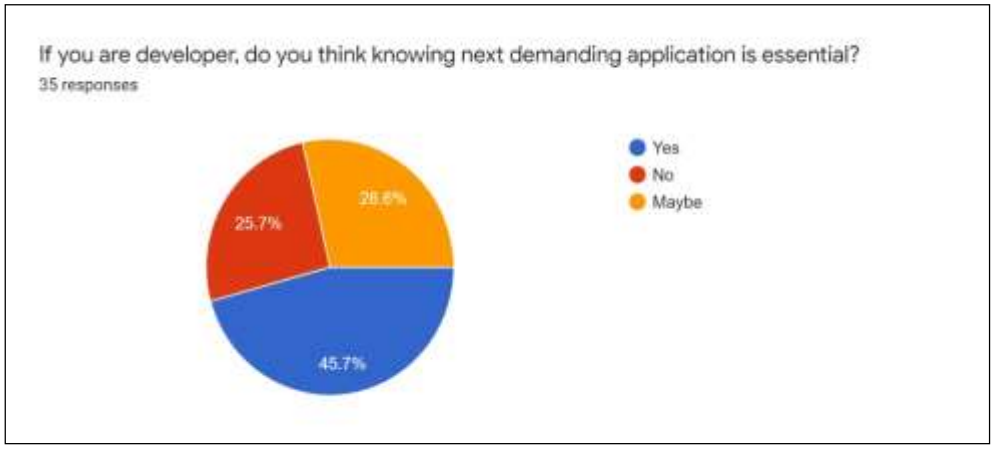

Figure 1. The pie chart above is to evaluate if the next demanding application is essential to the developer.

Based on the results, 45.7% of the users have responded that knowing the next demanding application is essential. Another 28.6 % of the users have responded 'Maybe' for the question and 25.7% of the respondents responded 'No' to the question. However, the result of 'Yes' remained significantly higher than the result 'No'.

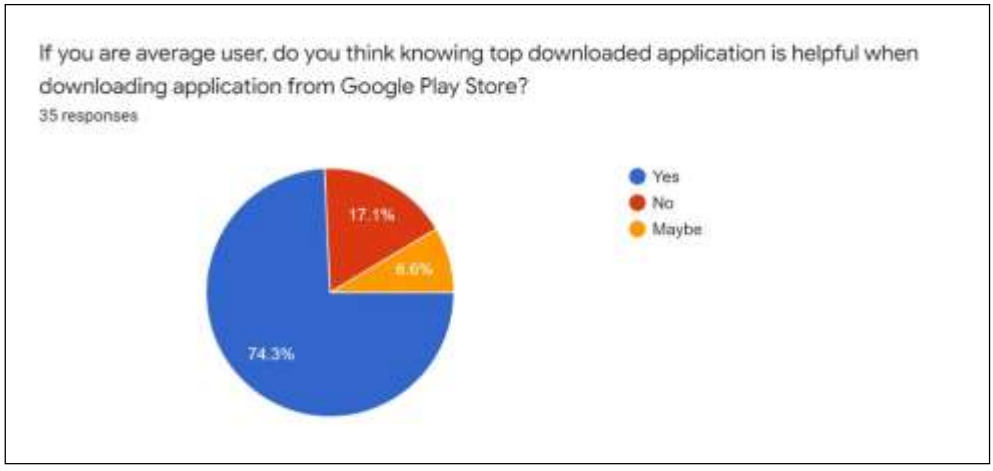

Figure 2. The pie chart above is to evaluate if knowing the top downloaded application is helpful to a normal user when downloading applications from the Google Play Store.

Based on the results, 74.3% of the respondents have agreed that knowing the top downloaded application is helpful when downloading applications from the Google play store. Another 17.1% of the respondents chose 'No' for the question, while 8.6% of the respondents responded 'Maybe' for the question. However, the result of 'Yes' remained significantly higher than for the result of 'No'.

Graphical User Interface Design Basic Operation of the System

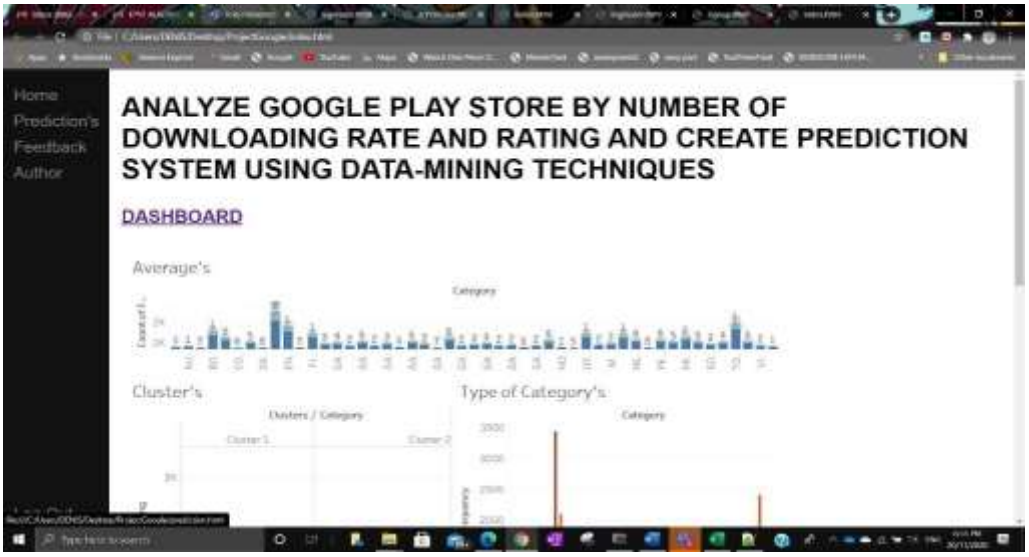

Figure 3. Home page of the Prediction System Developed to Analyze Google Play Store using Data- Mining Techniques.

The system displays the basic interface design for the web-based prediction system for Google Play Store which is developed by using PHP. This system will allow users to navigate through the homepage, prediction page, feedback page, author page and log out. This system also has the user sign-up function where users can register themselves and sign in function. The author has included a simple interface design where users can view a simple dashboard about Google Play Store analysis, which is about descriptive analytics from the data collected from Google Play Store.

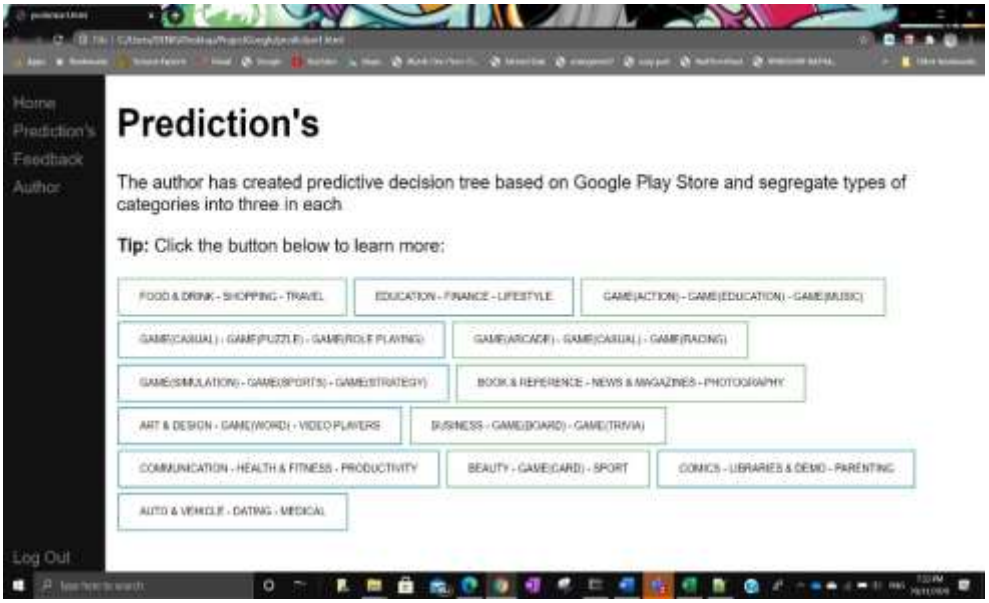

Figure 4. Prediction page for User Category selection

The prediction page is where the author has grouped 39 genres on the category of application types into 13 groups, which consist of 3 genres each. Users can view the predictions developed by the author using the decision tree to compare the outcomes and a brief explanation about the predictions of the user.

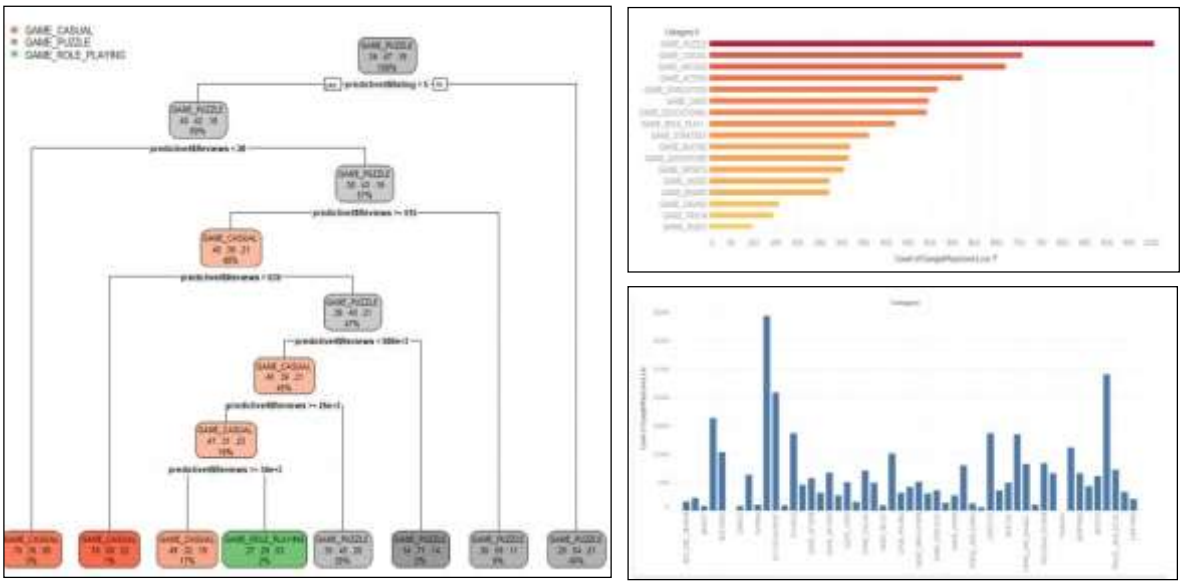

Figure 5. Descriptive and Decision Tree based on Prediction of the Google Play Store

The charts above provide information on descriptive and decision trees based on the prediction analysis. For the descriptive analysis, the author used Tableau. It is the most suitable tool to provide analysis of the data on business intelligence. It has the feature of creating a wide range of diverse visualization to show the data of the insights. This tool allows the user to drill the data to look at the impact in the visualization format which is easy for other users to understand. The decision tree developed using R-Studio, the tree is constructed via an algorithmic approach that identifies ways to split a data set based on various conditions. Each node from the tree will

split according to the probability of the popular genre. The descriptive and prediction analysis, through the R programming and Tableau, will benefit the users and bring progress to the company. Based on the results, it will help the company in the process of making decisions fast and accurately.

4. Conclusion

The proposed system described in this paper has been successfully designed and tested by the author. The system allows users, which caters to ordinary users and application developers, to identify the current demanding application and upcoming demanding application. Applications from the categories such as travel and local, education, game-action, game-puzzle, game-simulation, book and reference, parenting, dating, and medical have the highest probability of being upcoming demanding applications. Furthermore, knowing this will also elevate efficiency and productivity in application in the developing sectors.

References

The best Android apps (December 2019). (2019, December 9). Digital Trends. https://www.digitaltrends.com/mobile/best-android-apps/

Overview of Google Play services |. (n.d.). Google Developers. Retrieved November 9, 2021, from https://developers.google.com/android/guides/overview

Overview of Google Play services |. (n.d.). Google Developers. Retrieved November 9, 2021, from https://developers.google.com/android/guides/overview

SimilarWeb. (n.d.). Top Google Play Apps - Most Popular Apps in United States | All | Top Free.
https://www.similarweb.com/apps/trends/google/store-rank/us/all/top-free/

Google Play most popular app categories 2020. (n.d.). Statista. https://www.statista.com/statistics/279286/google-play-android-appcategories/

Wikipedia contributors. (2021, November 9). Google Play. Wikipedia. https://en.wikipedia.org/wiki/Google_Play