Pranali Jadav

Data Science Project

Best Category Selection for Apps on Google Play Store

Overview:

As part of the mobile app development class, we had created and released the Android application Voxiter and there is a possibility that the app will be maintained even after the end of class as a personal endeavor. Hence, I wanted to find out if there is a way to select an appropriate Category and Price for the app that will ensure biggest audience.

I found the dataset on Google Play Store apps for a certain year on Kaggle and decided to analyze it find for correlation between categories and a model that provides optimal parameters for best results.

Problem Specification:

The intend for the project is, given a list of possible categories it can be classified as, to select the best one that yields higher installs and/or higher profits. This will be achieved by analyzing the number of installs, Price type, and rating per category to see which one will be better. Since Google Play store provides a questionnaire to determine the content rating of an app, we cannot consider that as a user choice and thus limiting the user from making custom selection for the betterment of their app.

Data Selection:

The dataset is acquired from Kaggle, an online resource for dataset. The attributes include:

App: the name of the app

Category: the category the app belongs in.

Rating: the rating the app received by users.

Reviews: the number of reviews on the app

Size: the size of the app

Installs: the number of devices the app is installed on (greater than number)

Type: if the app is free or paid

Price: the price of app

Content Rating: the maturity rating of the app

Genres: the genres mentioned on the app

Last updated: the last update date for the app.

Current Ver: the version number of the app

Android Ver: the android OS version it requires to run.

Once imported, the dataset had to be changed since all fields were of type object. Thus, desired fields we type casted to appropriate types. Also, due to the lack of time and resources, certain fields were dropped that could potentially be more useful in determining the best selection.

Tools and Methodology:

To begin with, simple comparisons were created using numpy, panda, and matplotlib, on the empirical data to have a cursory look on the stats. To find the best categories to include in the model, I used the leaps library in R and found few categories that could be included. Though upon trying to find correlation between variables, there were no high R-squared values that signify correlation for any of the combination for the variables. This is was further confirmed when an ANOVA test was done on some of the categories, with number of Installs as the response variable, and the p-value was extremely low. Thus, it was evident that there couldn’t be an easier model to explain for the rise in the number of installs.

I then tried to use a data classification model from the Naïve Bayes to look for any models that might fit the data that I have inputted. And unfortunately, I could not find any models with a test accuracy of more than 22%. Perhaps adding the previously dropped fields might have added more of a pattern to the data for a model to fit.

Hence, I concluded with just analyzing the empirical data under different categories and values to determine if I can guide the developer in an appropriate direction to selecting the best category for marketing purposes.

Tools Questionnaire:

There are many other options that could have been explored and the data could be modified to be more flexible with the analyses, but due to time and resource constraint, I was not able to explore them. But it seems the dataset needed to be better set up to determine any correlation and additional fields were necessary to make any generalizations on the data.

Summary:

For the mobile application in question for this project, it was determined that putting the Category under education and a paid app was a successful endeavor as the low number of installs were balanced by the revenue of It being a paid app. If absolutely deciding to make it free, then putting it under Games would be the most beneficial.

Github link: <https://github.com/PranaliJadav/DataScienceProject>