Google Play Store App rating Prediction

Soumit Kar
Simplilearn Master Program 2019
Mail : soumitkar.5@gmail.com

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. We have used araw data set of Google Play Store from the Kaggle website.This data set contains 13 different features that can be used for predicting whether an app will be successful or not using different features. This data set is scraped from the Google Play Store. This journal talks about different classifier models that we used for prediction purposes and finding which one gives the highest accuracy. This journal also gives detailed information on feature extraction and the complete Data visualization done on this data set.

Datasets

The dataset taken is of Google play store application and is taken from Kaggle , which is the world‟s largest community for data scientists to explore ,analyze and share data. This dataset is for Web scratched information of 10k Play Store applications to analyze the market of android. Here it is a downloaded dataset which a user can use to examine the Android market of different use of classifications music, camera etc. With the assistance of this, client can predict see whether any given application will get lower or higher rating level. This dataset can be moreover used for future references for the proposal of any application. Additionally, the disconnected dataset is picked so as to choose the estimate exactly as online data gets revived all around a great part of the time. With the assistance of this dataset I will examine various qualities like rating, free or paid

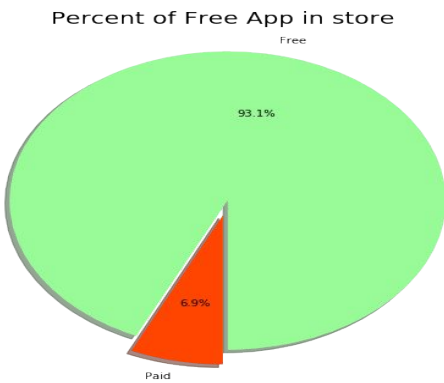
Objectives

The main goal of this project is to analyze different attributes of given application like application name, category, rating, reviews, size, installs, type, price, content rating, genres, last updated, current version, android version. And to find out the most rated and most reviewed apps and also to distinguish between the apps which are either free or paid

Python

Most of the info scientist use python due to the good built-in library functions and therefore the decent community. Python now has 70,000 libraries. Python is simplest programing language to select up compared to other language. That's the most reason data scientists use python more often, for machine learning and data processing data analyst want to use some language which is straightforward to use. That's one among the most reasons to use python. Specifically, for data scientist the foremost popular data inbuilt open source library is named panda. As we've seen earlier in our previous assignment once we got to plot scatter plot, heat maps, graphs, 3-dimensional data python built-in library comes very helpful.

Exploratory Data Analysis



Here we can see that 92.6% apps are free and 7.38% apps are paid on Google Play Store, so we can say that Most of the apps are free on Google Play Store.

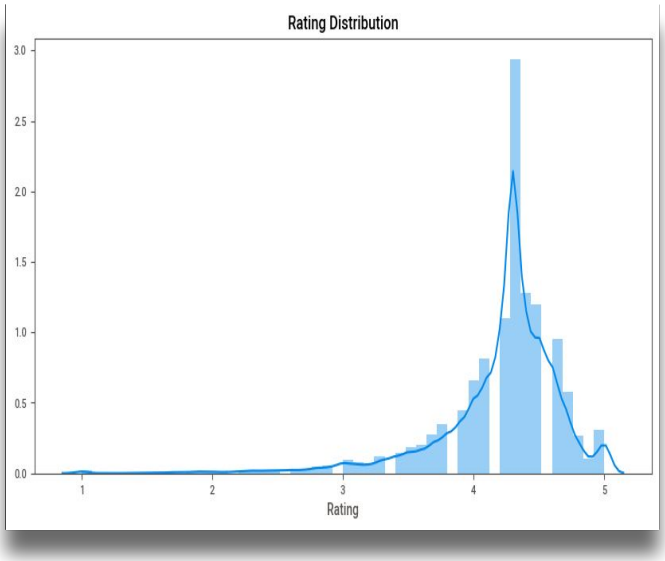
Missing Values

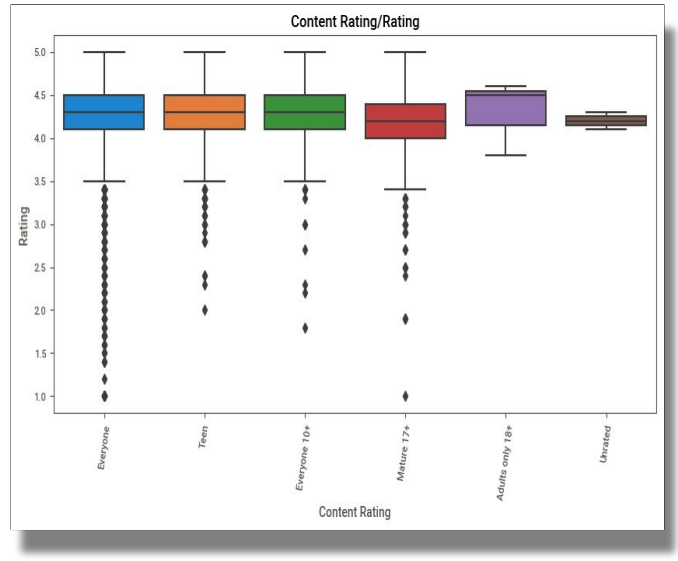
There are 1474 missing values in rating columns filled with median values.

Cleaning and Feature Engineering

Cleaning unwanted syntax, symbols. Put numeric values to Price column in place of 'M', 'K'. Remove reviews counts that greater than installs count

```
In [13]: GPS[GPS['Size']== 'Varies with divices']=np.nan

In [14]: GPS['Size']=GPS.Size.str.replace('M','e3')
         #convert K value
         GPS['Size']=GPS.Size.str.replace('K','e0')

In [15]: GPS['Size']=pd.to_numeric(GPS['Size'],errors='coerce')
         #replace Nan value with mean
         GPS['Size']=GPS['Size'].fillna(GPS['Size'].mean())

In [16]: GPS['Installs']=GPS.Installs.str.replace('+','')
         GPS['Installs']=GPS.Installs.str.replace(',','')
         GPS['Installs']=GPS['Installs'].astype(int)#convert to numeric

In [17]: GPS['Price']=GPS.Price.str.replace('$','')

In [18]: GPS['Price']=GPS['Price'].astype(float)

In [19]: GPS['Reviews']=GPS['Reviews'].astype(int)
         #remove reviews greater than installs
         GPS.drop(GPS[GPS['Reviews']>GPS['Installs']].index,axis=0,inplace=True)
```



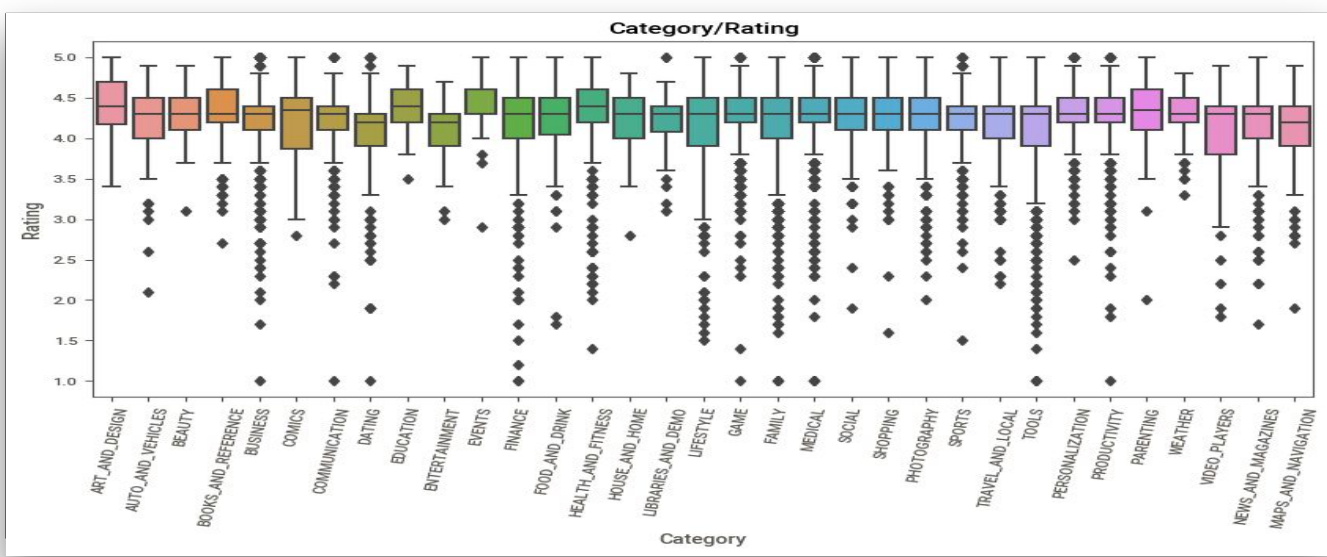


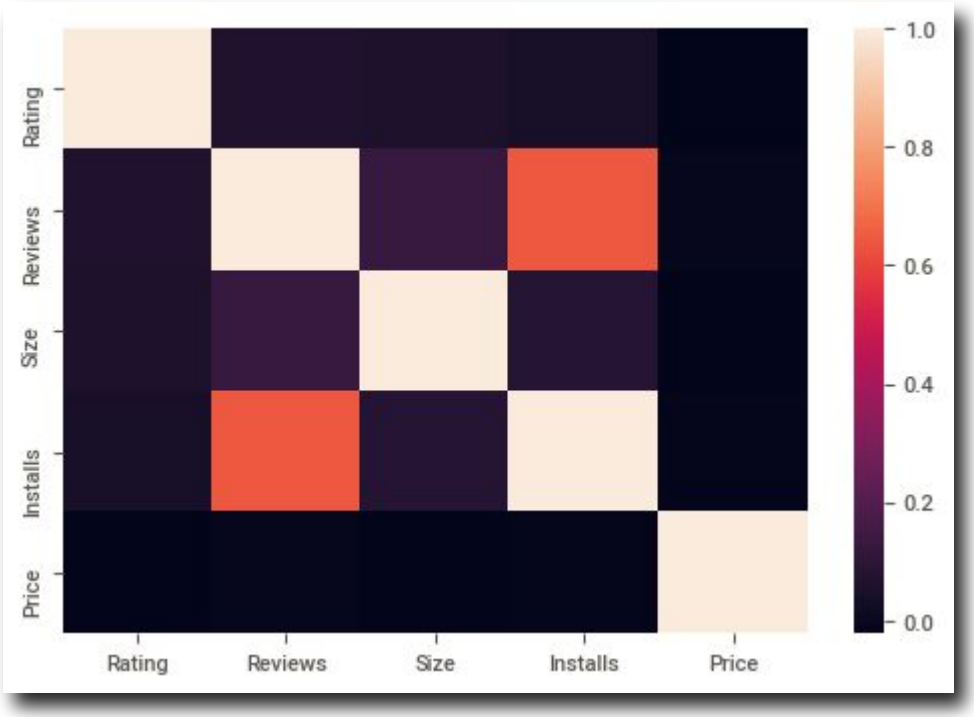
Rating data distributions

Many apps are rated between 3.5-5

Rating as per age Catagories

Adults(+18) Rated 3.7 - 4.5





Heat map

Transformation

Transfer categorical features by dummies and numerical variable to Label Encoder Methods

Train Data

Split train and test data to 75:25 ratio

Supervised Learning

Linear Regression Linear

Regression model is also used to find the variable importance of different variables with ratings. Although linear regression model is a simple regression model, it sometimes produces better results than other complex models. In this model, only numeric values are used. Therefore, when this model is applied to the dataset, only the numeric variables are considered.

K Nearest Neighbors

KNN is easiest supervised machine learning algorithm. It’s the foremost basic machine learning algorithm you'll find on scikit-learn. We will use KNN solve complicated problems. With the assistance of KNN we will do pattern recognition and data processing. KNN defines the similarity. From the given dataset KNN finds common groups between attributes. We split the info into training and test set. Then we will see what proportion similarity it becomes on the result.

Random Forest

Random forest regression is applied to all the variables the results of random forest determine the importance of all the variable and their influence on the rating. The results of random forest regression are evaluated using Mean Square Error. Random forest model is the first model that is applied to the dataset and the results of Random forest classification are computed for a number of variables to find the importance of these variables