Capstone Project-1 Submission

Play Store App Review Analysis

Prajakta Dangale, Aniket Deshmukh, Mahesh Landage

Data science trainees,

AlmaBetter, Bangalore

----------***----------

Prajakta Dangale- p.dangale7249@gmail.com

Aniket Deshmukh- deshmukhaniket013@gmail.com

Mahesh Landage- landge180@gmail.com

GitHub Link:

Prajakta Dangale- https://github.com/Prajakta1828D/Capstone-Project-Play-Store-App-Review-Analysis-Prajakta-Dangale

Aniket Deshmukh- https://github.com/aniket-deshmukh-data/Capstone-Project-Play-Store-App-Review-Analysis-Aniket-Deshmukh

Mahesh Landage- https://github.com/Mahesh-Landge/Play-Store-App-Review-Analysis---Capstone-Project-1-Mahesh-Landge

Abstract :

Google play store is engulfed with a few thousands of new applications regularly with a progressively huge number of designers working freely or on the other hand in a group to make them successful, with the enormous challenge from everywhere throughout the globe. Since most Play Store applications are free, the income model is very obscure and inaccessible regarding how the in-application buys, adverts and memberships add to the achievement of an application. In this way, an application's prosperity is normally dictated by the quantity of installation of the application and the client appraisals that it has gotten over its lifetime instead of the income is created. Application (App) ratings are feedback provided voluntarily by users and function important evaluation criteria for apps. However, these ratings can often be biased due to insufficient or missing votes. Additionally, significant differences are observed between numeric ratings and user reviews. This Study aims to predict the ratings of Google Play Store apps using machine learning Algorithms. I have tried to perform Data Analysis and prediction into the Google Play store application dataset that I have collected from Kaggle. Using Machine Learning Algorithms, I have tried to discover the relationships among various attributes present in my dataset such as which application is free or paid, about the user reviews, rating of the application.

Key Words: Google Play Store Apps, Ratings Prediction, Exploratory Data Analysis, Machine Learning.

Problem Statement:

Data is taken from the Google play store dataset. Every row contains various entries regarding a certain app. We will be doing Exploratory data analysis on this data set, which is a very important step in data science cycle, as it not only helps in taking very initial business decisions but also in preparing the data for further modelling for use in machine learning algorithms. Our objective will be to structure the data, clean it and present certain trends that we observe that can help us draw very preliminary conclusions about the probability of success of a newly launched app.

Introduction:

In this project we examine the different attributes present in the data set that affect the popularity of the application. We focused on to answer the questions like, what makes an app popular, what should be the price and size of the app, is there some trends in user sentiments.

In our data set we have two csv files for data analysis: Play Store data User Reviews At first, we analysis the play store data and in the play store data we have 10841 rows and 13 columns & in the user review data we have 64295 rows and 5 columns of data. We have to take the maximum outcomes from the data which help us to analysis the which type of app is most preferable and comparisons between different insights.

Our goal is to filter and make plots accordingly for a better EDA with respect to the final data. We need to explore and analyse the data to discover key factors

responsible for app engagement and success.

Data Transforming:

From the information of data frame, we can see that all the columns except rating have the object data type but some of the columns like, reviews, size, installs and price have the numerical value. So, we have to transform them in proper data type and also remove the unwanted values from the numerical columns like '+' and ',' from installs and '$' from price. In the size column we have some values in KB and some values in MB, so we transform all the values in MB.

Gathering data:

This step is about getting to know the data and understanding what has to be done before the data becomes useful in a particular context. This can be done by reading the CSV file and doing initial statistical analysis.

Though the dataset may seem to have the correct datatypes for each column, we need to check it. Inconsistent datatypes will create issues while dealing with problems.

The data set contains the following columns:

- **App:** This Column contains the name of the app
- **Category:** This contains the category to which the app belongs. The category column contains 33 unique values.
- **Rating:** This column contains the average value of the individual rating the app has received on the play store. Individual rating values can vary between 0 to 5.
- **Reviews:** This column contains the number of people that have given their feedback for the app.
- **Size:** This column contains the size of the app i.e. The memory space that the app occupies on the device after installation.
- **Installs:** This column indicates the number of time that the app has been downloaded from the play store, these are approximate values and not absolute values.
- **Type:** This column contains only two values- free and paid. They indicate whether the user must pay money to install the app on their device or not.
- **Price:** For paid apps this column contains the price of the app, for free apps it contains the value 0.
- **Content Rating:** It indicates the targeted audience of the app and their age group.
- **Genre:** This column contains to which genre the app belongs to, genre can be considered as a sub division of Category.
- **Last updated:** This column contains the info about the date on which the last update for the app was launched.
- **Current version:** Contains information about the current version of the app available on the play store.

- **Android version:** Contains information about the version of the android OS on which the app can be installed.

User Review Dataset:

User reviews data frame has 64295 rows and 5 columns. The 5 columns are identified as follows:

- **App:** Contains the name of the app with a short description (optional).
- **Translated Review:** It contains the English translation of the review dropped by the user of the app.
- **Sentiment:** It gives the attitude/emotion of the writer. It can be 'Positive', 'Negative', or 'Neutral'.
- **Sentiment Polarity:** It gives the polarity of the review. Its range is [-1,1], where 1 means 'Positive statement' and -1 means a 'Negative statement'.
- **Sentiment Subjectivity:** This value gives how close a reviewer's opinion is to the opinion of the general public. Its range is [0,1]. Higher the subjectivity, closer is the reviewer's opinion to the opinion of the general public, and lower subjectivity indicates the review is more of a factual information.

Data Cleaning and Preparation:

Our data set contains a large number of null values in the rating column, so we drop them. Some of the columns have a smaller number of null values, so we replace the null values in these columns with the mode value of that particular column. Our data set also contain the duplicate rows for a single application.

We also drop the duplicate rows because the rows contain the identical data. Also drop the rows, which have rating greater than 5. It is observed that some entries in the columns (e.g., Installs and Price) have some non-numeric characters, such as + and $.

Hence, it is necessary to clean our data as this may hinder the future computations.

- **Step1**: We write a function play store info (), that will display 5 attributes about all the columns: Data type, Count of non-null values, Count of null values, number of unique values in that column and percentage of null value in that columns in the play store dataset.
- **Step2**: we start off with the column 'Type' we can see that it has one null value. We checked this row and found out from the play store that it is a free app. We use filllna() function of the pandas library to fill this value
- **Step 3**: We drop the columns 'Current Ver', 'Android Ver' and 'last updated' from our dataset using the drop() function of the pandas library.
- **Step 4**: We can see that the 'Rating' column has 1474 null values. Due to low variations in the rating values and a lot of repeated values the 'median'

would be a suitable statistical indicator to replace the null values with. We calculate the mode of the column using the median () aggregate method, and fill this value in place of null values using the fillna() function.

- **Step 5:** We can see that the 'Reviews' column despite being a numerical indicator is of the 'object' data type, we will convert this to 'int' data type using the as type(int) function.
- **Step 6:** We can see that the size column, which should be numeric, is of the data type 'object', it also has characters 'k' and 'M' in the values which stand for kilobytes and Megabytes, we will replace the 'k' with 1000 and 'M' with 1000000. Some values also have '+' sign in them, which will be removed. Next, we will convert this column into 'int' datatype.
- **Step 7:** The 'Installs' column values contain the characters '+' and ',' which are going to prevent us from converting this column into a numeric datatype. We will get rid of these using the strip() and replace() functions.
- **Step 8:** The values in the column 'Price' might have the '$' sign in some values and the column is of the datatype 'object'. We will first remove the '$' sign using the **strip ()** function and then convert the column into 'int' datatype.
- **Step 9:** Handling the duplicates in the App column we drop the no of duplicate rows that are present in the App columns.
- **Step 10:** We write a function Ur info (), that will display 5 attributes about all the columns: Data type, Count of non-null values, Count of null values, number of unique values in that column and percentage of null value in that columns in the User review dataset.
- **Step11:** In the User review dataset the columns are App, Translated Review, Sentiment, Sentiment Polarity, Sentiment Subjectivity in this total 26863 NaN value are present so we drop them using dropna () function.

EXPLORATORY DATA ANALYSIS:

Exploratory Data Analysis (EDA) is an approach to analysing datasets to summarize their main characteristics, often with visual methods. EDA is used for seeing what the data can tell us before the modelling task. It is not easy to look at a column of numbers or a whole spreadsheet and determine important characteristics of the data. It may be tedious, boring, and/or overwhelming to derive insights by looking at plain numbers. Exploratory data analysis techniques have been devised as an aid in this situation.

1. Problem Statement - We shall brainstorm and understand the given data set. We shall study the attributes present in it and try to do a philosophical analysis about their meaning and importance for this problem.

- **Free vs Paid**

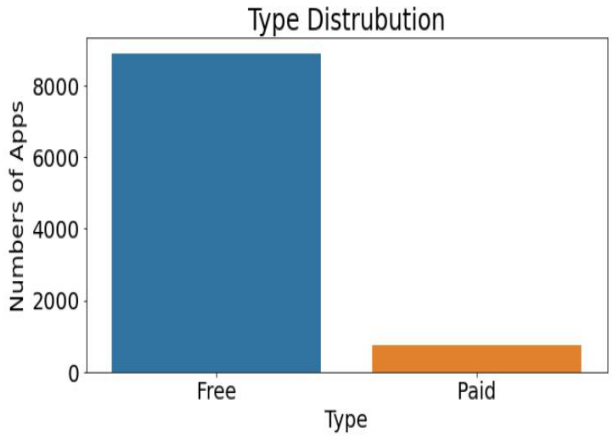

Fig -1: Free vs Paid

Around majority app are freely available on play store.

- **Apps rating**

In the below plot, we plotted the Apps rating

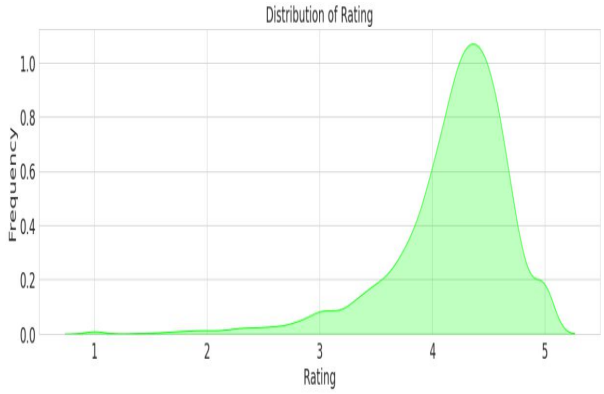

Fig -2: Distribution of App rating

From this distribution plotting, it implies that most of the apps in the Play Store are having rating higher than 4 or in the range of 4 to 4.7

- **Content Rating**

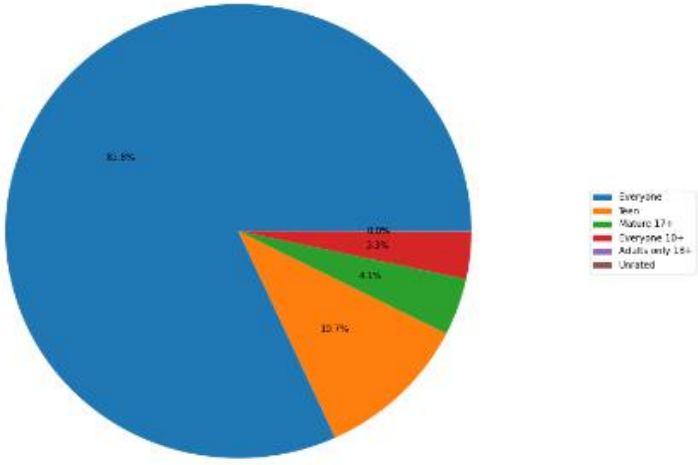

Fig-3: Content Rating

- From the above plot we can see that Everyone category having majority of apps count.
- A majority of the apps (81.80%) in the play store are can be used by everyone. The remaining apps have various age restrictions to use it.

- **Numbers of installs vs apps category**

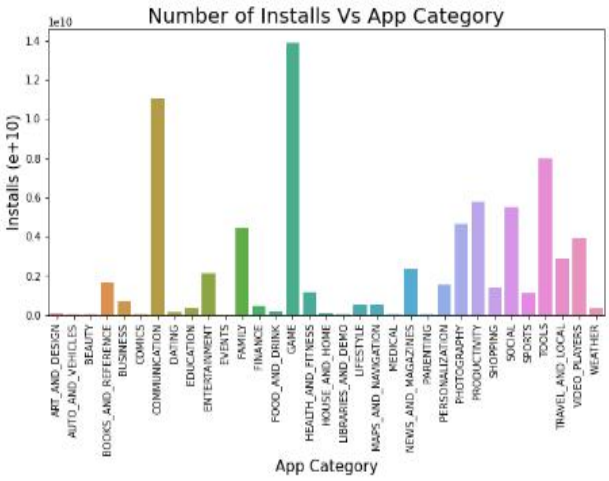

Fig-4: Numbers of installs vs apps category

- Gaming has maximum number of installed app.
- On second Position communication have maximum number of apps installed.

- **Co-Relation in Merged Dataset.**

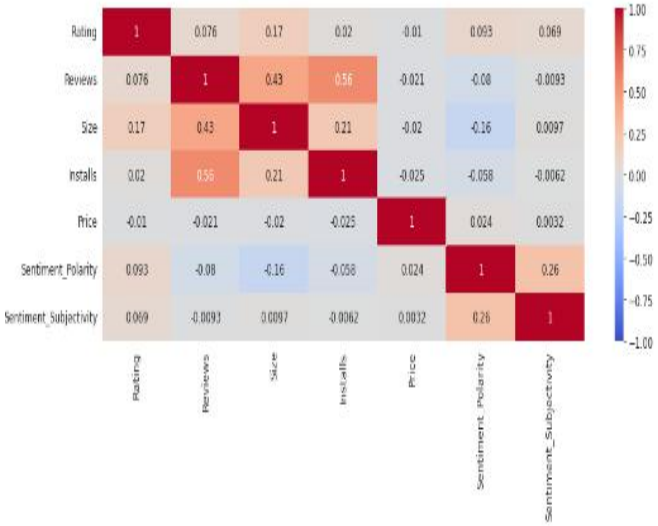

Fig-5: Co-Relation in Merged Dataset.

- In this correlation matrix, There is not a significant relationship between Rating, Reviews, Size and Installs with respect to the Sentiment polarity and Sentiment subjectivity.

- **Relation between Rating, Size and Type of app**

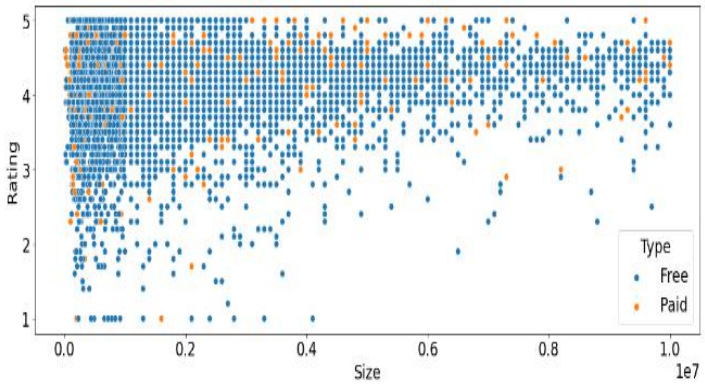

Fig-6: Relation between Rating, Size and Type of app

- From this scatter plot, we can imply that majority of the free apps are small in size and having high rating. While for paid apps, we have quite equal distribution in term on size and rating.

- **Top 10 Genres of App**

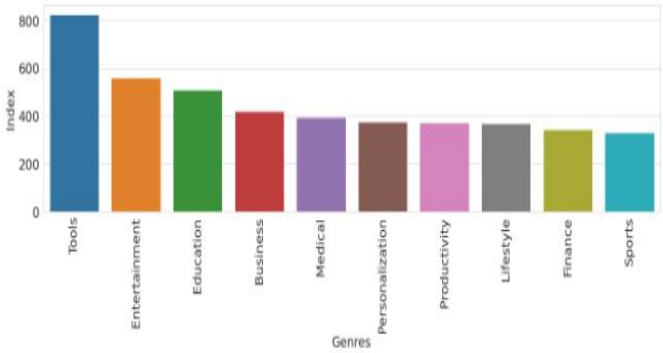

Fig-7: Top 10 Genres of App

- Top three Genres are Tools, Entertainment & Education

- **User Sentiment Analysis**

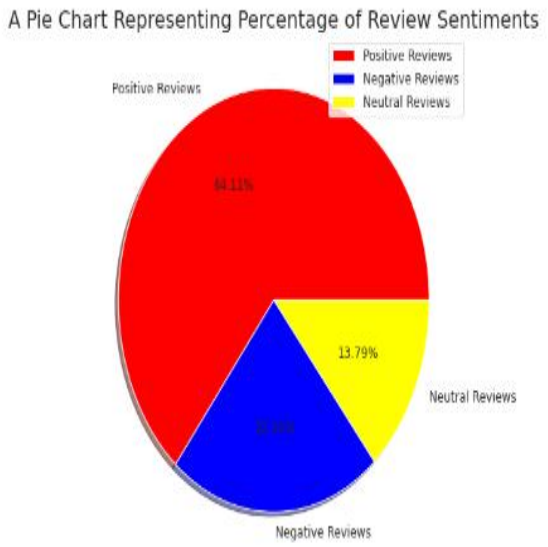

Fig-8: User Sentiment Analysis

- From Sentiment column, 64% are Positive, 22% are Negative and 14% are Neutral Reviews.
- Positive Sentiment have larger percentage, it has 3 times more than negative sentiment.

Challenges Faced:

1. Reading the dataset and comprehending the problem statement. Our major challenge was data cleaning.

2. Handling the error, duplicate and NaN values in the dataset.

3. 13.60% of reviews were NaN values, and even after merging both the dataframes, we could not infer much in order to fill them. Thus, we had to drop them.

4. There is so much more which can be explored. Like we have current version, android version available which can be explored in detail and we can come out with more analysis where we can tell how does these things effect and needs to be kept in mind while developing app for the users.

Conclusion:

The dataset contains possibilities to deliver insights to understand customer demands better and thus help developers to popularize the product. Dataset can also be used to look whether the original ratings of the app matches the predicted rating to know whether the app is performing better or worse compared to other apps on the Play Store.

Other than that, the charts shown above actually implies that most of the apps having good ratings of above 4.0 are mostly confirmed to have high number of reviews and user installs. There are some spikes in term of size and price but it shouldn't reflect that apps with high rating are mostly big in size and pricy as by looking at the graphs they are most probably are due to some minority.

- Category with the highest number of installs: Game.
- Category with the highest average app installs: Communication.
- Most popular app in the Play Store based on the number of reviews: Facebook.

- The apps whose size varies with device has the highest number average app installs.
- Category in which the paid apps have the highest average installation fee: Finance.
- There are 20 free apps that have been installed over a billion times.
- There are top 10 most installed app.
- We understand by the graph most expensive top 10 app.
- Instagram, WhatsApp Messenger, Messenger – Text and Video Chat for Free, Subway Surfers, Facebook these are the top 5 reviewed app.
- Most competitive category: Family.

References:

- Stackoverflow
- GeeksforGeeks
- Analytics Vidhya
- Python libraries documentation
- Researchgate.net
- https://www.academia.edu