Google Play Store Apps Project Proposal
By Xu Han

I. SUMMARY

While many public data sets provided Apple store data, there is not enough analysis into apps in Google Play store, partly because it's more difficult to scrape data from Google Play store.

Nowadays, as mobile phones become more popular, people tend to spend more time on their phones and app usage increases a log. While this is a great opportunity for many app developers, it also becomes a challenge for many developers/businesses to develop a popular app.

In this project, I will try to predict the number of installs (target variable) from some features of the app itself, shown in **googleaplystore**. I am trying to find out what kind of apps are more popular and tend to stay longer in people's phones. Since the exact number of installs was not available, but an estimate was given, it is treated as categorical variable, so the problem will be a classification problem.

II. DATA

The Google Play Store Apps was downloaded from Kaggle (Link). The information was scraped from the Google and the dataset was updated eight months ago.

The dataset contains two files: **googleaplystore** and **googleplaystore_user_reviews**. The first file has 10,841 entries, each representing an app listed in Google Store. There are 13 features in the original dataset.
These features include:

App: the name of the App.

Category: the category the App belongs to.
Its values include: 1.9, ART_AND_DESIGN, AUTO_AND_VEHICLES, BEAUTY, BOOKS_AND_REFERENCE, BUSINESS, COMICS, COMMUNICATION, DATING, EDUCATION, ENTERTAINMENT, EVENTS, FAMILY, FINANCE, FOOD_AND_DRINK, GAME, HEALTH_AND_FITNESS, HOUSE_AND_HOME, LIBRARIES_AND_DEMO, LIFESTYLE, MAPS_AND_NAVIGATION, MEDICAL, NEWS_AND_MAGAZINES, PARENTING, PERSONALIZATION, PHOTOGRAPHY, PRODUCTIVITY, SHOPPING, SOCIAL, SPORTS, TOOLS, TRAVEL_AND_LOCAL, VIDEO_PLAYERS, WEATHER.

Rating: numerical value represented using decimal numbers. Ranging from 1 to 5, with 5 being the highest, usually round off to one decimal place such as 4.1.

Reviews: number of reviews this app has received.

Size: file size of the app, measured in megabyte, unit symbol M.

Installs: number of installs, categorical variable, target variable. 0, 1+, 5+, 10+, 50+, 100+, 500+, 1000+, 5000+, 10_000+, 50_000+, 100_000+, 500_000+, 5_000_000, 10_000_000+, 50_000_000, 100_000_000+.

Type: categorial variable, Paid or Free.

Price: Price of the app, 0 if free, measured in US dollar.

Content Rating: categorical variable representing the audient this app is suitable for. Its values include: Mature 17+, Everyone, Teen, Everyone 10+ etc.

Genres: categorical variable, similar to Category. its values include Action, Adventure, Action & Adventure, Arcade, Art & Design, Auto & Vehicles, Beauty, Board, Books & Reference, Brain Games, Business, Card, Casino, Casual, Comics, Communication, Creativity, Dating, Education, Educational, Entertainment, Events, Finance, Food & Drink, Health & Fitness, House & Home, Libraries & Demo, Lifestyle, Maps & Navigation, Medical, Music, Music & Audio, Music & Video, News & Magazines, Parenting, Personalization, Photography, Pretend Play, Productivity, Puzzle, Racing, Role Playing, Shopping, Simulation, Social, Sports, Strategy, Tools, Travel & Local, Trivia, Video Players & Editors, Weather, Word.

Last Updated: last date this app was updated, when this dataset was made.

Current Ver: current version of the app.

Android Ver: Android operating system requirement.

The second file **googleplaystore_user_reviews** contains the user reviews of some apps listed in the previous file. Each row represents one user review. There are 64295 reviews and 5 columns. These columns include:

App: the name of the App.
Translated _Review: review in English.
Sentiment: positive, neutral, negative, and nan values.
Sentiment_polarity: a value between 0 and 1, with one denoting positive and 0 denoting negative, related to sentiment.

Sentiment_subjectivity: a value between 0 and 1.

III. PUBLIC WORK

This dataset has been used to predict App Ratings or perform sentiment analysis. For the App Rating prediction, the author used columns from **googleaplystore** to perform linear regression, SVR, and random forest models to predict app rating, which he treated as a continues variable. For sentiment analysis, all columns from the **googleplaystore_user_reviews** have been used to count words and create word cloud.

IV. PREPROCESSING

StandardScaler is used to transform reviews, which is the numbers of reviews each app has received, and this number can be very large.
OneHotEncoder is used to transform Category and Genres, both of which are categorical variables and do not have an order.
LabelEncoder is used to transform the target variable: number of installs. In the original dataset, the exact number was not published. Instead, a range was given, such as 100+, 500+, 10,000+. It does have order, that's why I treat it as a categorical variable and use LabelEncoder to transform it.