Midterm Project Report

Xiongfeng Wang

Introduction

This project will perform a supervised machine learning process on the Google play store Apps dataset[1] using rating as the target variable. This is a regression problem considering that the rating scales from 1 to 5, decile between integers.

This is an interesting task because it allows us to predict the rating of an App with certain features. For the existing Apps, comparing the prediction result with the true value can assist in finding outliers with specialty. For recently developed Apps, a precise prediction result reveals the feedback in the future which saves the value of time, or gives people a hint about the App's customer experience.

The dataset initially has 10841 samples with 1 key (App), 1 target variable (Rating) and 11 features. For description and property, please refer to the table below.

There are plenty of public projects on this dataset, mostly focusing on EDA, data visualization, machine learning on rating prediction and data analysis. The top voted notebook[4] analyzed almost every column of the dataset except for the version ones and discussed some relationship between certain columns, for example, Users prefer to pay for apps that are light-weighted. The notebook's conclusion is neat but market-related practical. A trending notebook[5] plots the data with different plotting packages but just like most other visualizations, author discussed the relationship between rating and one feature individually. The top voted machine learning notebook[2] provides integer encoding RFR model, dummy encoding SVR model and random forest regressor model but cannot conclude which model has the best predictive.

EDA

Caption:

The Distribution of Category plot shows the that samples in Family and Game category is way higher than others, later in the Distribution of Category and Subcategory[3] we will find out the reason for this is that there are plenty of subcategories under these two groups.

One problem with the data set is that feature ‘Genre’ actually stands for the subcategory of the App if applicable. The Distribution of Category and Subcategory plot clearly reveals the component of the main category. As we can see for Family and Game, there are more subcategories in it.

The Scatter Matrix for float & int variables plot shows the relationship between variables with the type of float & int. If we look at the histogram for Distribution of Size, Price and Days, which are not included here, we will find out the varibable is quite imbalance, x-axis must be semiloged to show a

fair plot. This explains why some of the matrix plot has heavy skewness. This suggests that we should apply StandardScalerEncoding later.
I use bin = 39 for the Rating over Category graph because of the rating scale I mentioned above, we noticed that Family and Gaming seems to have a higher average score than the overall rating.

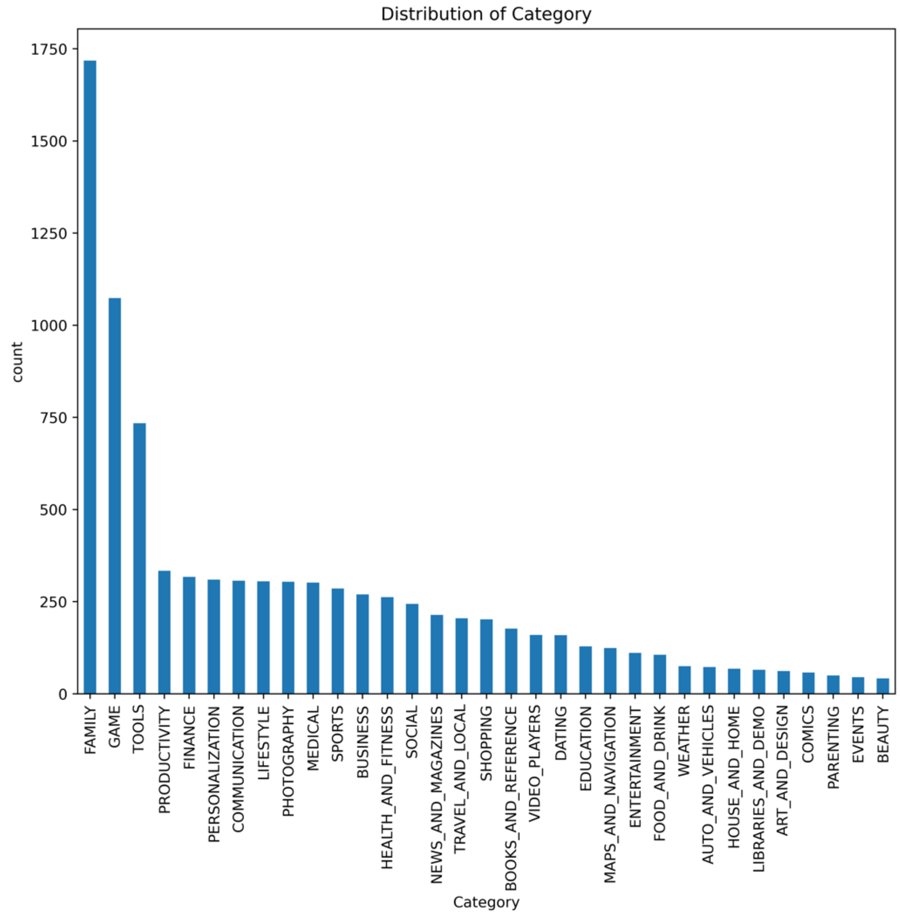
Data Processing

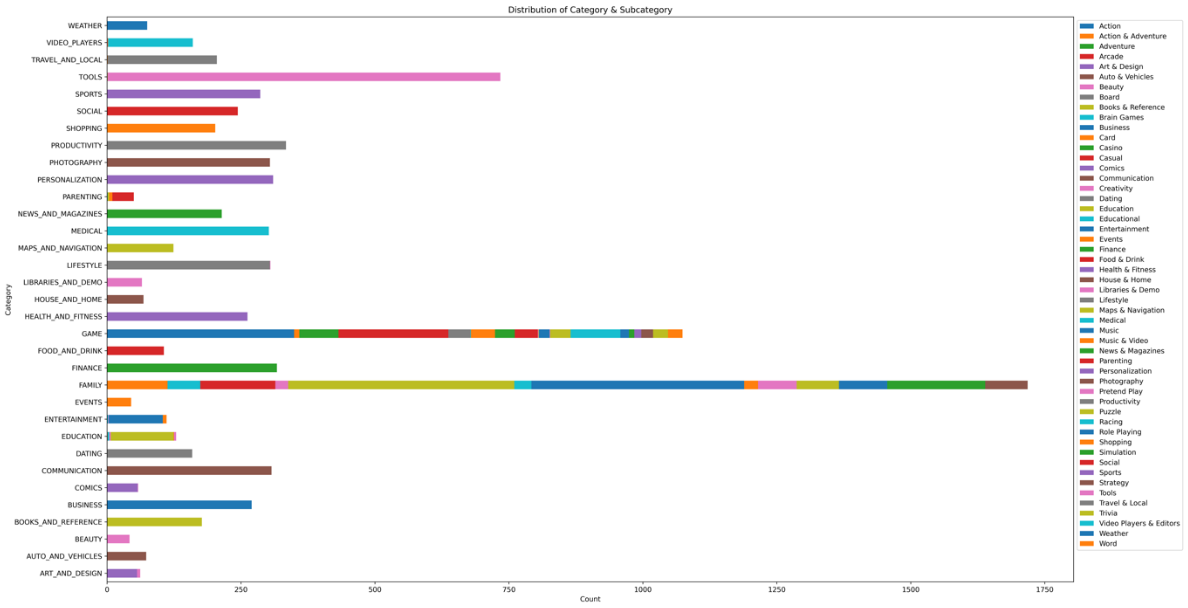
This is not an IID dataset because clearly there are group structures in the dataset: Category and Subcategory. Also this is not a time series dataset because even though for very small amount of Apps, there are samples for different version of it and the rating might vary, the proportion is too small to be consider as a general feature. Most of the Apps with different version only differ in number of reviews by around 1%.
Considering there are 8893 samples with a valid target features, the size of the data set is fair to apply a 8-1-1 split. Category with lowest counts is 42 for Beauty, 9 categories' count below 100, and only 2 categories' count above 800, so the 8-1-1 split is rational because we don't need to worry about group in train/test set should not appear in the other. We can implement plenty of n_split on the dataset.
Since there's an obvious group structure, I will choose GroupShuffleSplit to split the dataset. I choose GroupShuffleSplit over GroupKFold because the data set is also imbalance as we can see from the distribution of categories, or some other features.
For Encoder selection and reason, please refer to the table below.

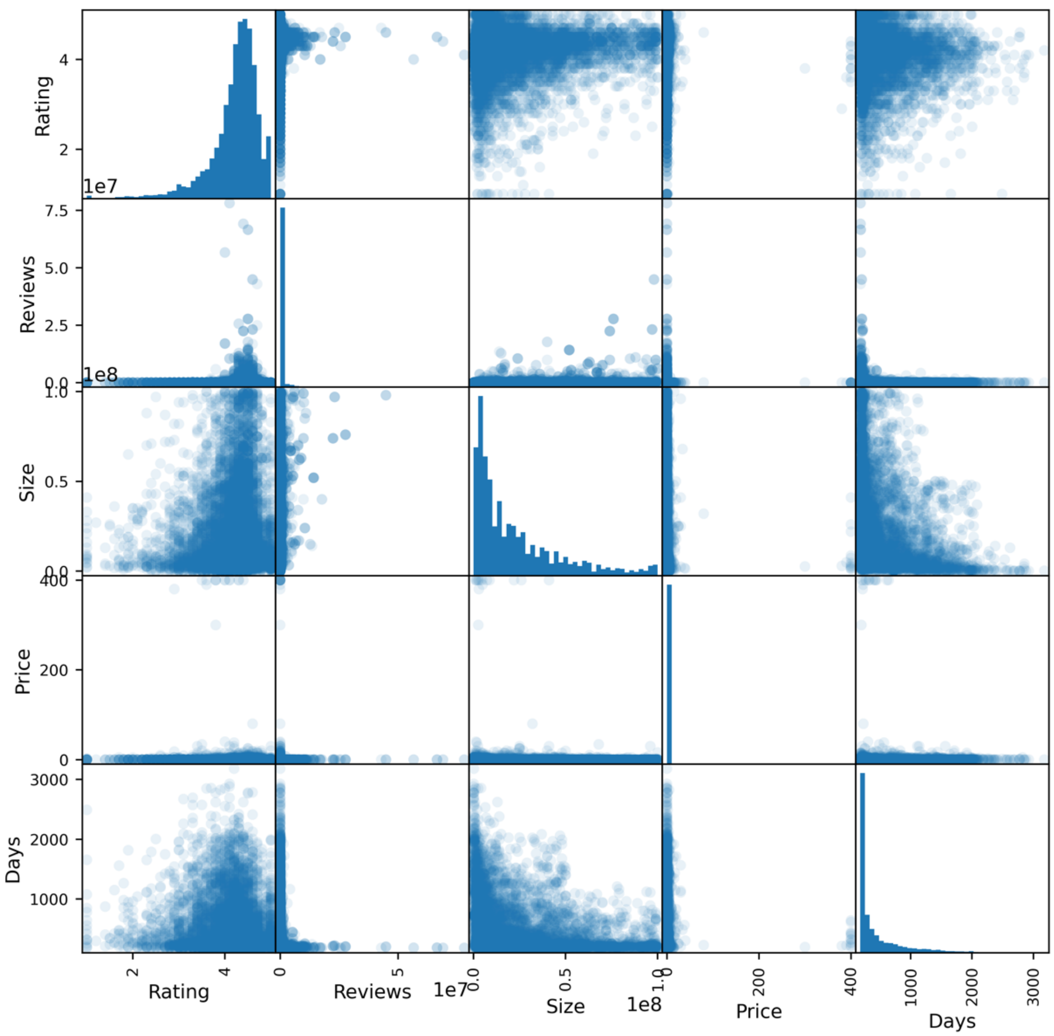
	Variable	Description		Encoding	Note on reasons
Key	App	Name of the App	Categorical	No Encoding	
Feature	Category	Category of the App	Categorical	OneHotEncoder	No order
	Reviews	Number of reviews	Numerical	StandardScaler	Skewness
	Size	Size of the App	Numerical	StandardScaler	Skewness
	Installs	Scale for number of installs	Categorical	OrdinalEncoder	With order
	Type	Free or Paid	Categorical	OneHotEncoder	No order
	Price	Price of the App	Numerical	StandardScaler	
	Content Rating	Restriction rating	Categorical	OneHotEncoder	No order
	Genres	Category; Subcategory	Categorical	N/A (see Subcategory)	
	Last Updated	Last updated date in Month DD, YYYY	Numerical	N/A (see Days)	

	Current Ver	Current verison of the App	Categorical	N/A	
	Android Ver	Android version that supports	Categorical	OneHotEncoder	No order
	Subcategory	Subcategory	Categorical	OneHotEncoder	No order
	Days	Day since last update	Numerical	StandardScaler	Skewness
Target	Rating	Rating for the App	Numerical	StandardScaler	

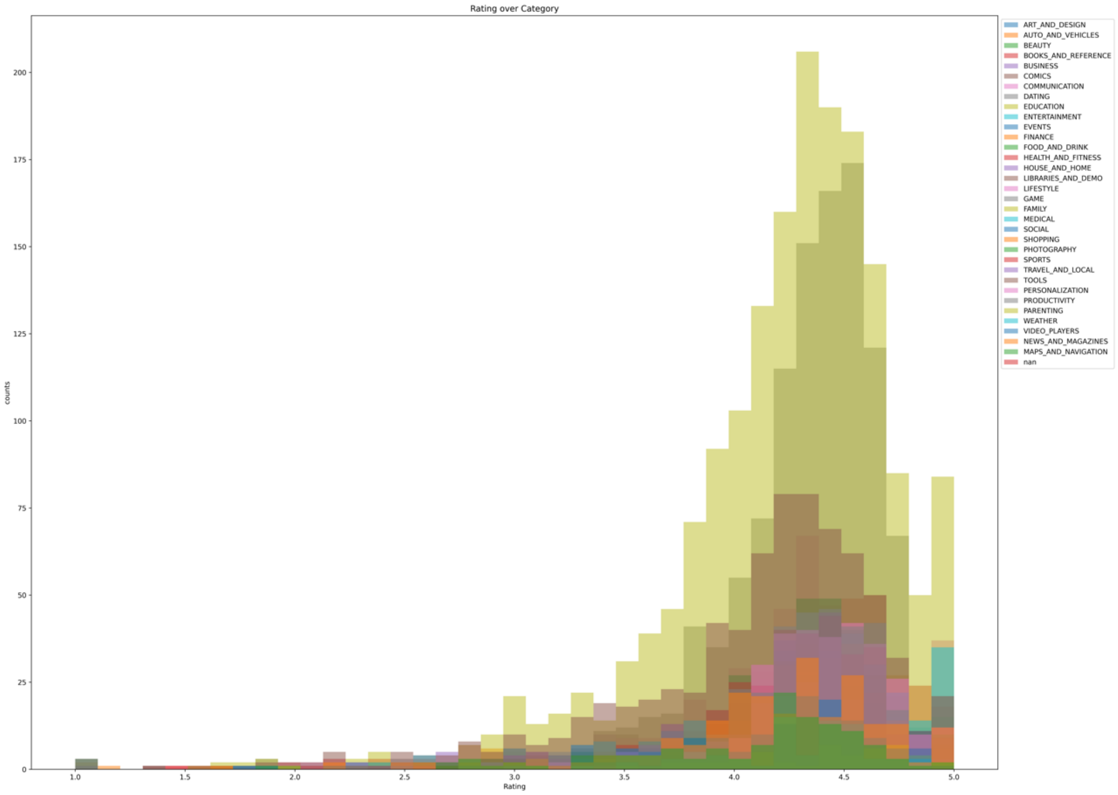
Figures


Distribution of Category
count
1750
1500
1250
1000
750
500
250
0
FAMILY
GAME
TOOLS
PRODUCTIVITY
FINANCE
PERSONALIZATION
COMMUNICATION
LIFESTYLE
PHOTOGRAPHY
MEDICAL
SPORTS
BUSINESS
HEALTH_AND_FITNESS
SOCIAL
NEWS_AND_MAGAZINES
TRAVEL_AND_LOCAL
SHOPPING
BOOKS_AND_REFERENCE
VIDEO_PLAYERS
DATING
EDUCATION
MAPS_AND_NAVIGATION
ENTERTAINMENT
FOOD_AND_DRINK
WEATHER
AUTO_AND_VEHICLES
HOUSE_AND_HOME
LIBRARIES_AND_DEMO
ART_AND_DESIGN
COMICS
PARENTING
EVENTS
BEAUTY
Category


Distribution of Category & Subcategory
Category
WEATHER
VIDEO_PLAYERS
TRAVEL_AND_LOCAL
TOOLS
SPORTS
SOCIAL
SHOPPING
PRODUCTIVITY
PHOTOGRAPHY
PERSONALIZATION
PARENTING
NEWS_AND_MAGAZINES
MEDICAL
MAPS_AND_NAVIGATION
LIFESTYLE
LIBRARIES_AND_DEMO
HOUSE_AND_HOME
HEALTH_AND_FITNESS
GAME
FOOD_AND_DRINK
FINANCE
FAMILY
EVENTS
ENTERTAINMENT
EDUCATION
DATING
COMMUNICATION
COMICS
BUSINESS
BOOKS_AND_REFERENCE
BEAUTY
AUTO_AND_VEHICLES
ART_AND_DESIGN
0
250
500
750
1000
1250
1500
1750
Count
Action
Action & Adventure
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 & 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



Scatter Matrix for float & int variables


Rating over Category
counts
0
25
50
75
100
125
150
175
200
Rating
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
ART_AND_DESIGN
AUTO_AND_VEHICLES
BEAUTY
BOOKS_AND_REFERENCE
BUSINESS
COMICS
COMMUNICATION
DATING
EDUCATION
ENTERTAINMENT
EVENTS
FINANCE
FOOD_AND_DRINK
HEALTH_AND_FITNESS
HOUSE_AND_HOME
LIBRARIES_AND_DEMO
LIFESTYLE
GAME
FAMILY
MEDICAL
SOCIAL
SHOPPING
PHOTOGRAPHY
SPORTS
TRAVEL_AND_LOCAL
TOOLS
PERSONALIZATION
PRODUCTIVITY
PARENTING
WEATHER
VIDEO_PLAYERS
NEWS_AND_MAGAZINES
MAPS_AND_NAVIGATION
nan

Reference

1. Gupta, L. (2019, February 03). Google Play Store Apps Dataset. Retrieved October 13, 2020, from https://www.kaggle.com/lava18/google-play-store-apps
2. Jemseow, J. (2018, September 27). Machine Learning to predict app ratings. Retrieved October 13, 2020, from https://www.kaggle.com/jemseow/machine-learning-to-predict-app-ratings
3. Siddiqi, S. (2018, September 28). Google Play Store Apps - Data Cleaning. Retrieved October 13, 2020, from https://www.kaggle.com/sabasiddiqi/google-play-store-apps-data-cleaning
4. Gupta, L. (2018, September 19). All that you need to know about the Android market. Retrieved October 13, 2020, from https://www.kaggle.com/lava18/all-that-you-need-to-know-about-the-android-market
5. Jha, R. K. (2020, January 27). ML to Visualization &amp; Prediction of App Ratings. Retrieved October 13, 2020, from https://www.kaggle.com/rajeshjnv/ml-to-visualization-prediction-of-app-ratings