Erdal Terkivatan

Google Play Store - Rating Analysis
Final Report

Problem Statement

Google Play, formerly Android Market, is a digital distribution service operated and developed by Google. The aim of this project is to provide insights to android application developers based on ratings and the number of installs.

This project will allow android application developers to create strategies that improve the application and increase the demand for the application. It will also help the companies to understand their position with respect to their competitors.

Data visualization techniques will be used to answer the following questions.

1. What is the most popular category?
2. What is the app with the largest size?
3. What are the apps that were downloaded most?
4. What is the app with a large number of reviews?

We believe that more questions will come up as we work on the data. We will also utilize a regression model to predict ratings.

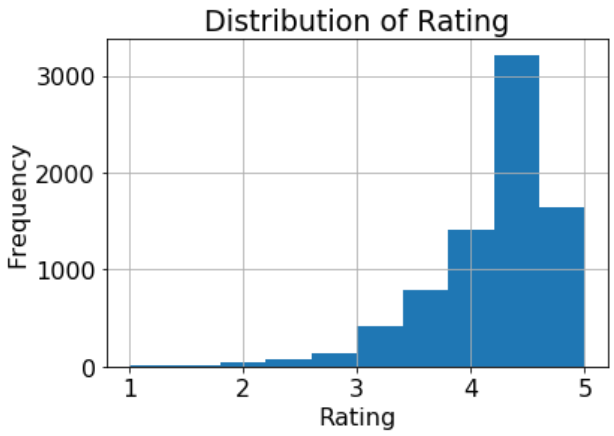
Data Wrangling

The data is scraped from the Google Play Store including the following features: rating, size, Installs, type, price, Content Rating, Category, Genres, Last Updated, Current Ver, Android Ver.

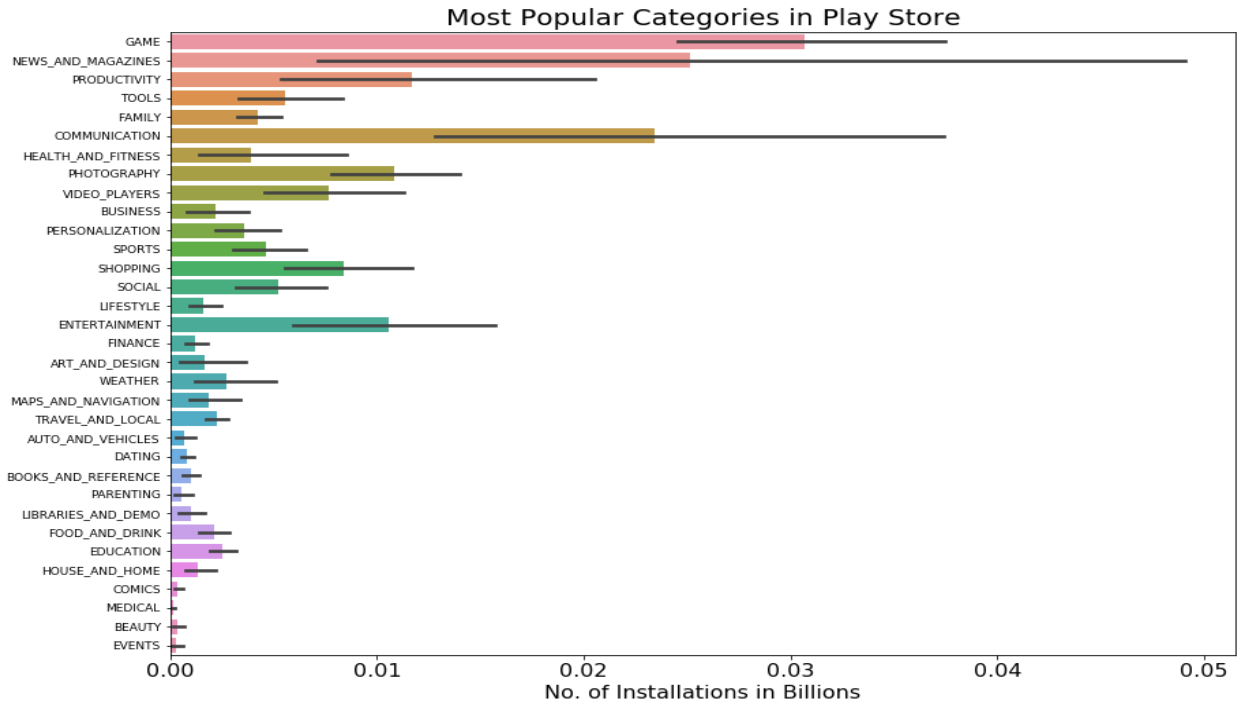
We dropped the genres variable as it has identical values as the “Category” variable. We also performed appropriate data conversions. The features “Current version” and “Android Version” are also dropped as they were irrelevant to the research objectives. Finally, non-ASCII characters and null values are removed from the data for the modeling part.

Exploratory Data Analysis

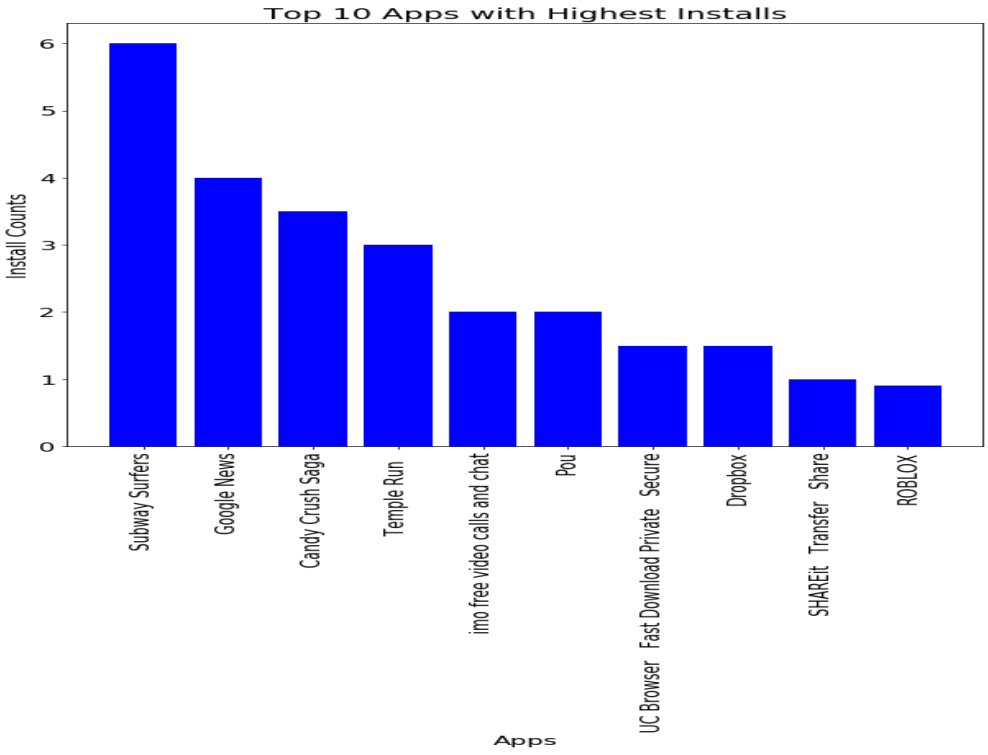
The histogram below displays the distribution of ‘Rating’. We can see that the distribution is left-skewed. In another word, there are some small rating values (2 and below) that cause the skewness in the data.



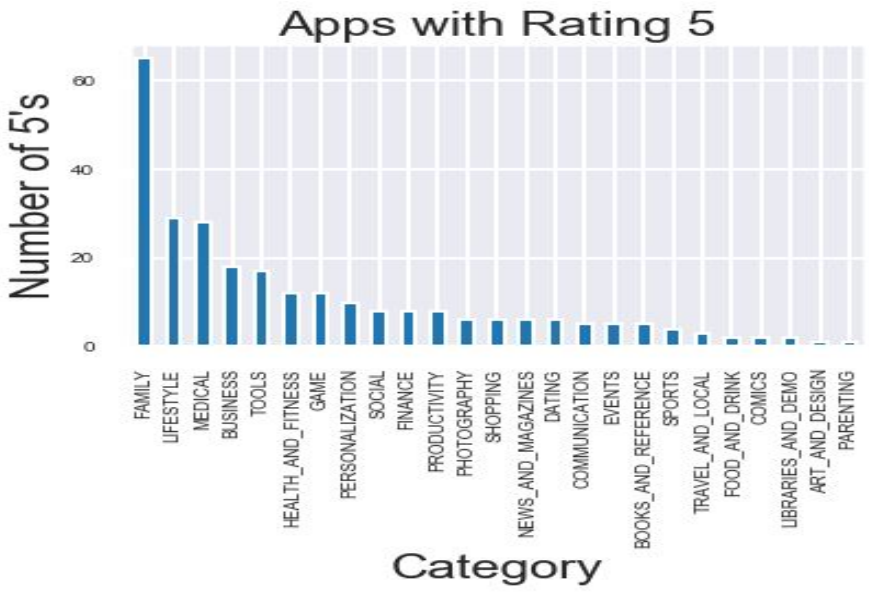
We wanted to see what category is most popular in the Google Play Store based on the number of installs. We found that the most popular category is the “Game” category.



Another question that we think that it is worth asking is " What are the top 10 Apps that have maximum installs? The following graph shows that the "Subway Surfers" in-game category is the most installed app followed by "Google News".



We compared the categories in terms of the number of times that it received a rating of 5. Based on the graph below, one can see that the "Family" category has the maximum number of rating 5.

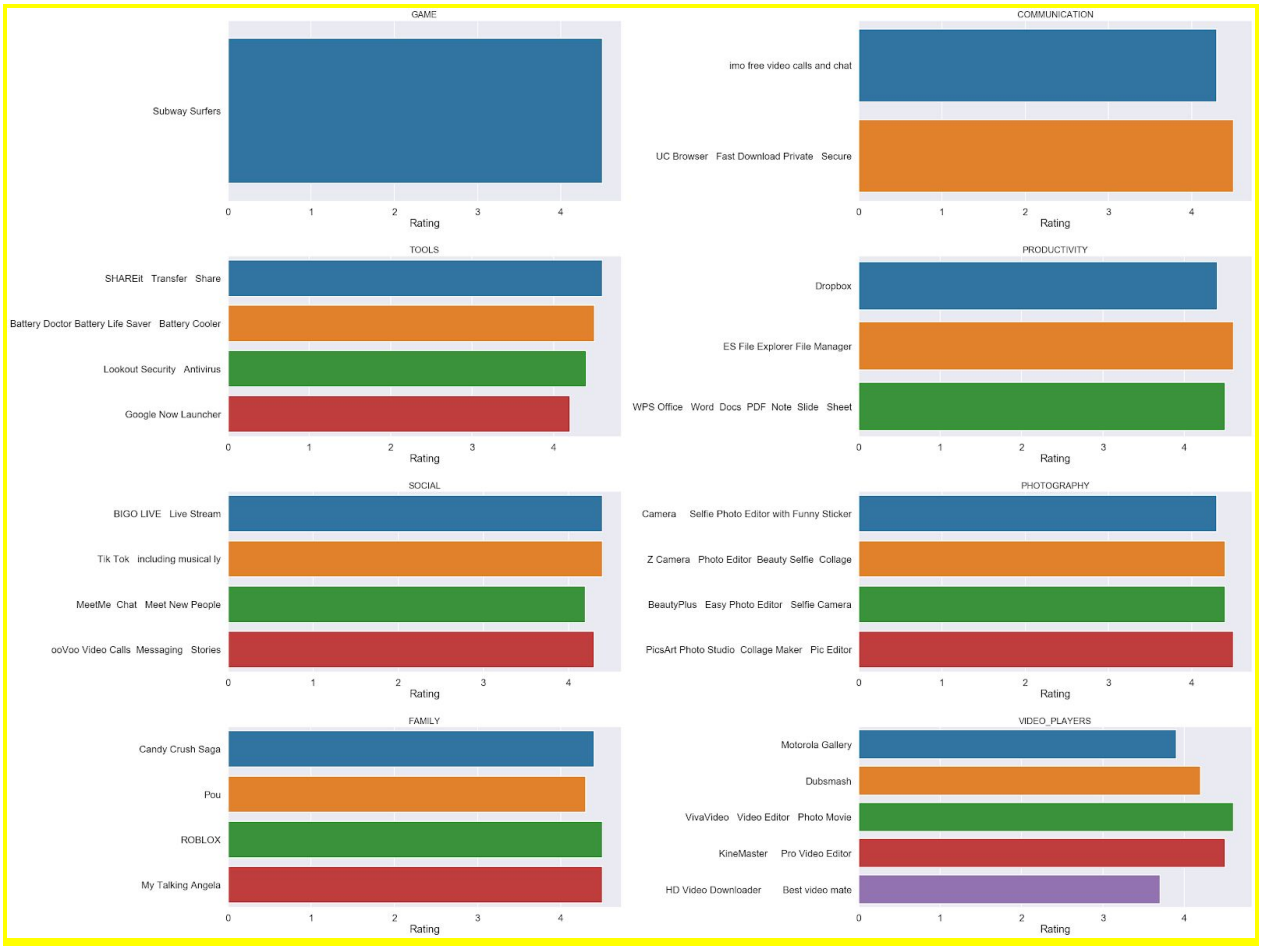


Apps reviewed most

The graphs above show the top five apps with the most reviews. Namely;
Apps with the most reviews in the following categories are as follows :

1. GAME - Subway Surfers
2. COMMUNICATION- UC Browser
3. TOOLS - Battery Doctor
4. PRODUCTIVITY - ES file Explorer
5. SOCIAL - TikTok
6. PHOTOGRAPHY- PicsArt
7. FAMILY- Candy Crush Saga
8. VIDEO PLAYERS - ViVa Video

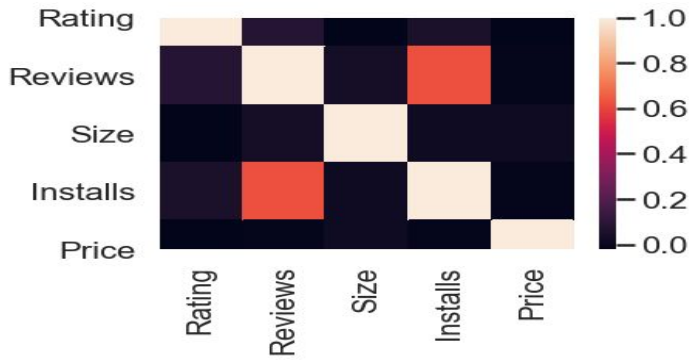
Apps with the highest rating



The graphs above display the apps with the highest rating in some of the categories. We can see that the top apps with the highest rating are as follows:

1. GAME - Subway Surfers
2. COMMUNICATION- UC Browser
3. TOOLS - SHAREit
4. PRODUCTIVITY - ES file explorer
5. SOCIAL - BIGO live Stream and TikTok
6. PHOTOGRAPHY- PicsArt
7. FAMILY- ROBLOX and My Talking Angela
8. VIDEO PLAYERS - ViVa Video

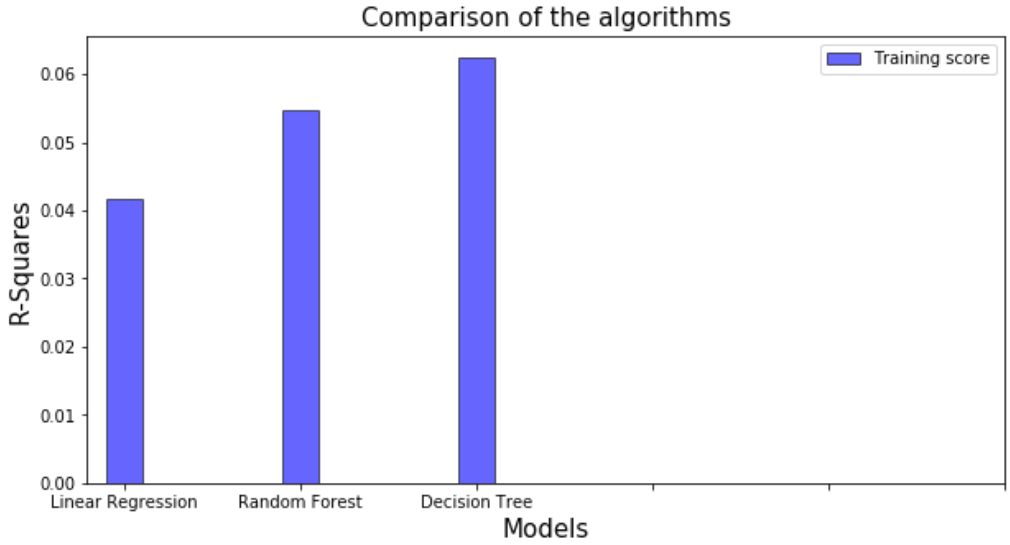
Finally, we performed a correlation analysis to see if there is a strong correlation between “Rating” and the other variables. The following heatmap displays the correlation between the variables.



We can see that the correlation between rating and the other variables are very weak. However, we observe that there is a moderate positive correlation between “ Installs” and “ Reviews” which is 0.63.

Model Selection

We used Random Forest Regressor, DecisionTree Regressor, and Linear Regression models to predict the variable “Rating”. To assess the model performance we focused on R-square values.



The bar graph above includes the R-square values for each model. We can see that linear regression is the worst performing model. On the other side, Decision Tree Regression is the best performing model. Therefore, we performed hyperparameter tuning on the decision tree regression model.

We used the grid search algorithm to find the optimal parameter values for the decision tree regression model. The R-square value after tuning the parameters is 6.7%. The R-square for the

decision tree algorithm was 6.3% before the grid search. Thus, we do not see significant improvement in the model.

Future Research

We see that all the regression algorithms we used achieved very low R-square values. The maximum R-square value, 6.7% belongs to the Decision Tree Regression model. The reason for poor performances is because our target variable “Rating” is very weakly correlated with the other variables. Therefore, we suggest that different variables should be considered to make informed predictions on the rating variable.

We also think that people should be encouraged to give reviews as “Reviews” is moderately positively correlated with “Installs”.