Google Play Store

App Popularity Analysis

Erdal Terkivatan

The Problem

What are the factors that affect app rating?

The Data

Model Variables

Target Variable: Rating

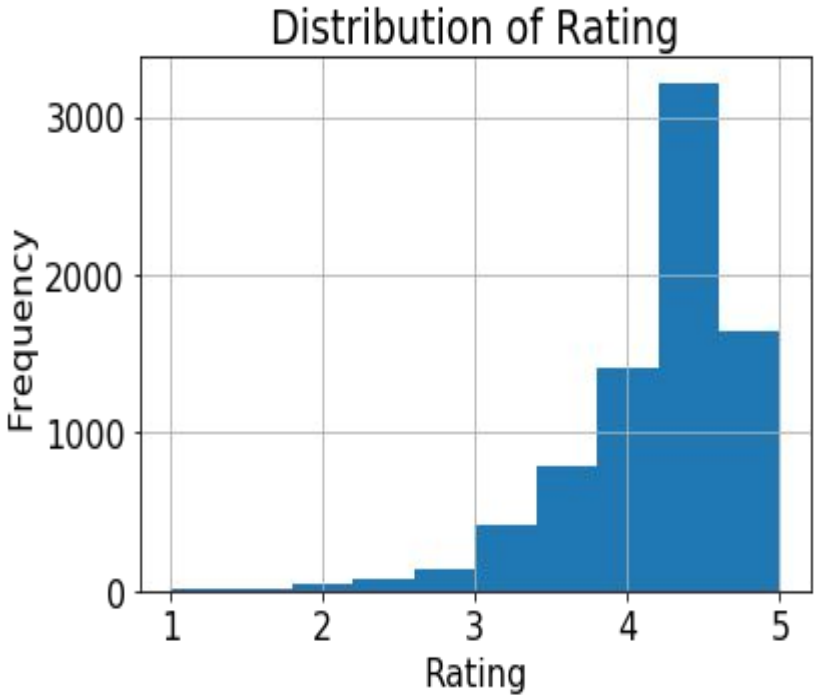
1. Category (Genres)
2. App Size
3. App Type (Free vs. Paid)
4. Number of Installs
5. Number of reviews
6. Age group that rated the app
7. App name

Data Wrangling

Activities performed on the data

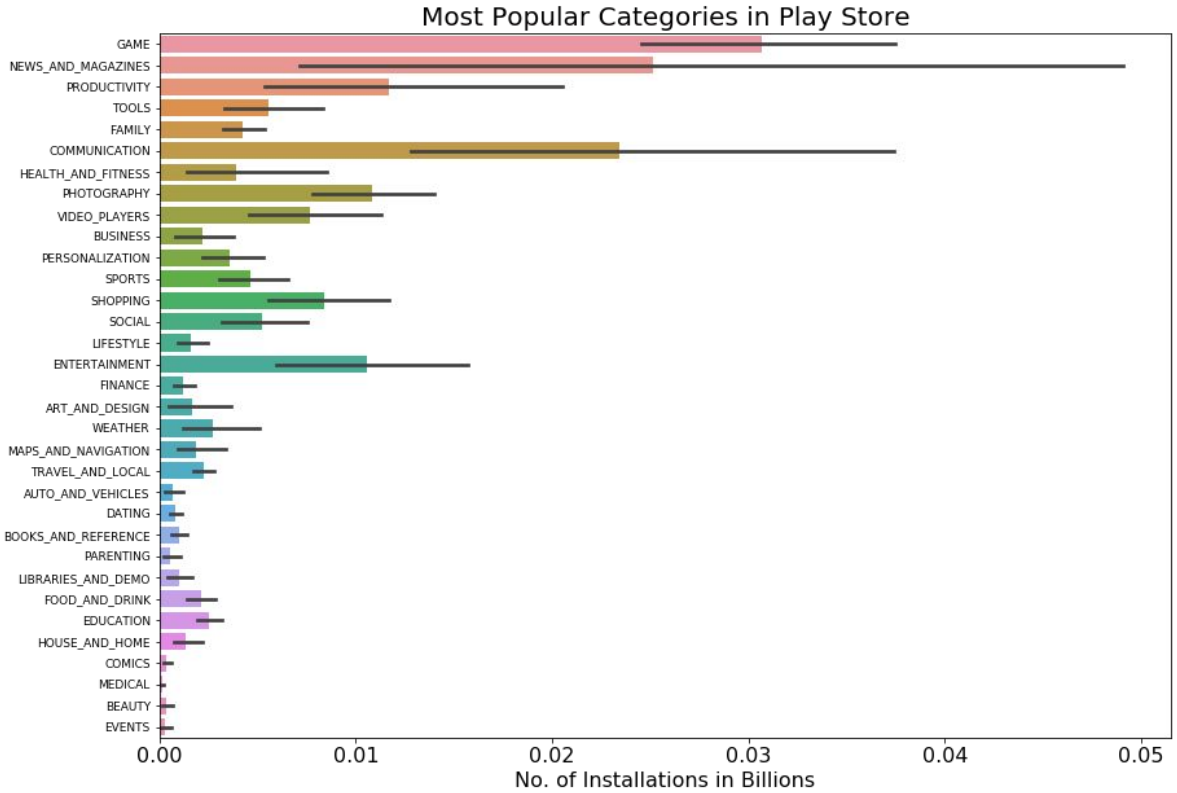
1. Data type conversion
2. Drop irrelevant columns such as current version, android version.
3. Remove all non-ASCII characters in the data
4. Apply appropriate filling method to null values
5. Detect duplicates

Exploratory Data Analysis



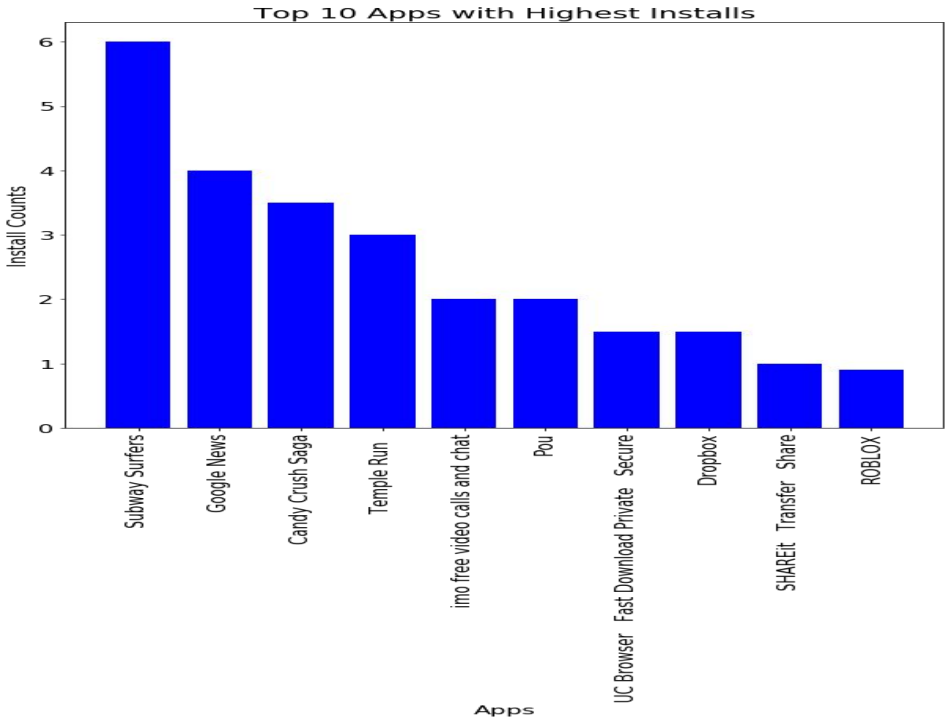
- The rating variable has left-skewed data due to the small values (2 or below)

Most Popular Category (Based on Installs)



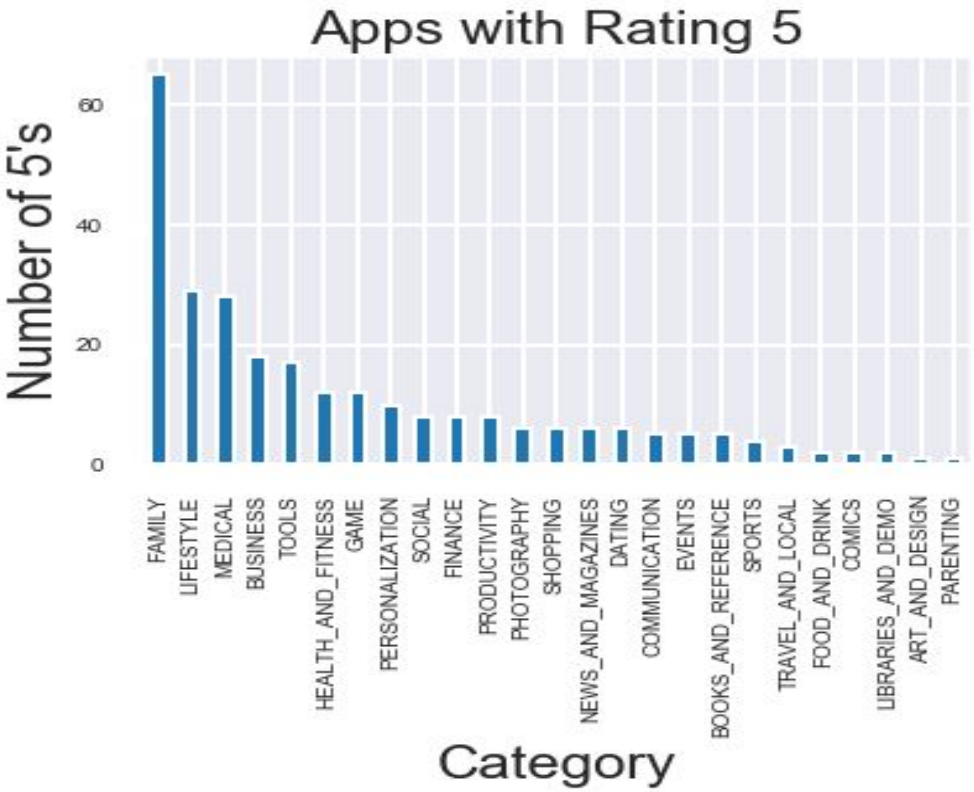
- Apps in game category has the maximum number of installs

Top 10 Apps installed most



- “Subway Surfers” is installed most of the time followed by “Google News” and “Candy Crash Saga”

What Category is rated as 5 most?



- “Family” category rated 5 most of the time.

Correlation Analysis

	Rating	Reviews	Size	Installs	Price
Rating	1.000	0.080	-0.019	0.053	-0.021
Reviews	0.080	1.000	0.037	0.626	-0.010
Size	-0.019	0.037	1.000	0.017	0.018
Installs	0.053	0.626	0.017	1.000	-0.011
Price	-0.021	-0.010	0.018	-0.011	1.000

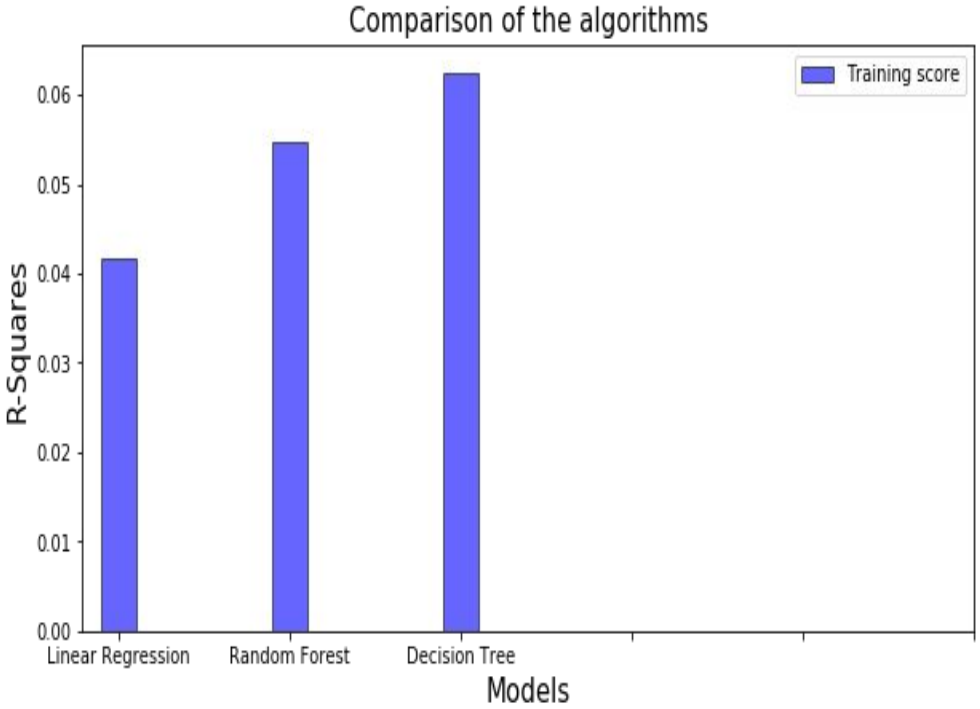
- Almost no correlation between rating and the other variables
- “Installs” and “ Reviews” are moderately positively correlated

Model Selection

Models

1. Random Forest Regression
2. Linear Regression
3. Decision Tree Regression

Performance Comparison



- Linear Regression has the lowest R-square
- Decision tree performs the best

Hyperparameter Tuning

- Applied hyperparameter tuning on decision tree regression algorithm
- The R-square increased from 6.3% to 6.7%

Conclusions

- None of the algorithms performed well .
- The highest R-square achieved by decision tree algorithm is 6.7%

- **The reason for poor performance:** The variables included in the model are almost not correlated with the response variable

Suggestion

- Consider different explanatory variables