Predicting Ratings and Analyzing Root Causes for Customer Sentiments Regarding Google Play Store Apps

Priscilla Lee, Sonali Mahendran, Rachel Protacio

Princeton University, Department of Computer Science, Princeton, NJ

Abstract

- Understanding customer feedback and being able to predict and adapt to market trends are essential to successful app development.
- In this work, we implement supervised and unsupervised techniques to explore Google Play Store App data:
  - (1) Supervised: Predicting ratings based on market data and review sentiments
  - (2) Unsupervised: Determining the root causes for users' positive and negative critiques of an app
- Analysis of app ratings reveals that an app receives high ratings if it is free, uses moderate memory, is well maintained, and boasts a large number of user reviews.
- Analysis of customer reviews reveals that users appreciate apps that are easy to use and install, pose helpful or captivating content, are inexpensive, and display ads minimally.

Linear Regression for Predicting App Ratings

Weights Assigned to Features

L_2 Penalized Ridge Regression

- Last Updated: April 2012 -0.2963
- Genres: Educational -0.2193
- Price -0.1954
- Number of Reviews +0.2115
- Last Updated: August 2018 +0.2697

L_1 Penalized LASSO Regression

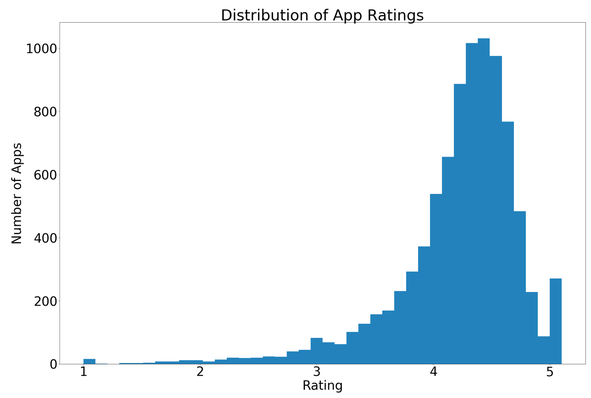
- Last Updated: April 2012 -2.5744
- Price -0.3168
- Genres: Educational -0.3051
- Last Updated: August 2018 +0.2689
- Number of Reviews +0.4352

Regression	Mean Squared Error
Ordinary Least Squares	0.255319
L_2 Penalized Ridge	**0.252109**
L_1 Penalized LASSO	0.252118

Background and Approach

Kaggle Google Play Store Apps Dataset

We performed our analysis on a web-scraped dataset for 10K Google Play Store apps [1], which includes features such as rating, category, number of reviews, number of installs, size, price, and date of last update.



Regression Methods for Predicting Ratings

App ratings are continuous, and we assume that they have a linear relationship with the features.

- Ordinary Least Squares Linear Regression
- L_2 Penalized Ridge Regression
- L_1 Penalized LASSO Regression

LDA Methods for Root Cause Analysis of User Reviews

In order to obtain topics from the free-response user reviews, LDA requires a bag of words representation of the text. We achieved this using the following approaches.

- Count Vectorizer
- Term Frequency - Inverse Document Frequency (TF-IDF)

Latent Dirichlet Allocation (LDA) for Review Analysis

5 Topics for Unique Negative Unigrams

- game, terrible, disappointed, difficult, load, slow
- worst, useless, stupid, tried, data, install
- bad, hate, wrong, game, people, makes
- annoying, fake, boring, stars, stop, don't
- horrible, pay, customer, money, google, awful

3 Topics for Unique Positive Unigrams

- best, ok, perfect, free, thing
- love, nice, awesome, easy, able
- good, great, amazing, better, think

5 Topics for Positive Bigrams

- i think, easy use, i need, i would, awesome app
- i love, love app, love it, i cant, thank you
- very good, it's good, i like, it's great, 5 stars
- the best, very nice, good app, works great, really good
- great app, it's ok, every time, best ever, i wish

5 Topics for Negative Bigrams

- worst ever, this game, very poor, takes forever, forever load
- i hate, bad bad, worst game, hate game, do not
- very bad, the worst, very annoying, even though, ads annoying
- every time, i try, this worst, get rid, i even
- i can't, can't even, please fix, waste time, keeps crashing

Reference

[1] Lavanya Gupta. Kaggle Google Playstore Apps Dataset.https://www.kaggle.com/lava18/google-play-store-apps. Accessed: 2019-04-17.