Getting 1 million + Installs on Google Play Store!

Team After 6: Angela Yoon, Ki Hyun, Tharun Mani Raj, Sara Mehta

Topic and Motivation

Relevant Background Information

In this project, we are investigating a dataset containing information on various aspects of **applications** on the **Google Play Store**, the most popular, growing App Store. There is immense literature on product and app development techniques that can be used to get the first 100,000 downloads for an application. This literature inspired our investigation to delve deeper into the **relation between application features** that are easier to grapple with, such as ratings and the price of an application, and the **number of installs**.

Primary Motivation

With all four of us team members being extremely interested in **application development** and some of us being Android users, we are interested in exploring the different factors of a given application on Play Store that can predict the app's number of downloads. Actionable insights can be drawn for developers to work on and **capture the Android market** based on our results. Moreover, these results will be insightful for **prediction** of future applications' number of downloads as well.

With the above motivations and dataset in mind, the general research question our team explored is, **"For Google Play Store Apps, to what extent are the aspects of the application such as ratings, genre, pricing, size, and other variables in the dataset predictive of the odds of the application having more than 1 million downloads?"**

Data

Source

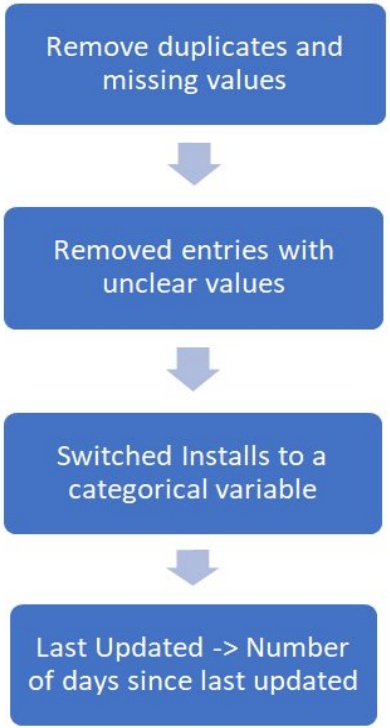
- This specific dataset was obtained from Kaggle. The data itself was scraped from the Google Play Store.
- The data was scraped on **February 3rd, 2019** in **NYC**.
- In regards to that, the scraping would be specific to the **geographical location** and might not include the same apps if the scraping procedure was carried out from another location.

Description

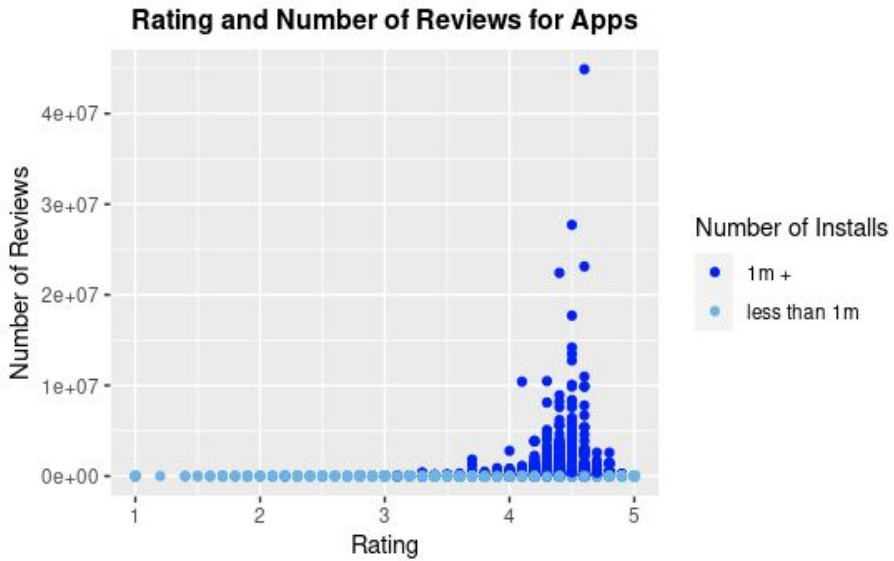
- The dataset contains 10841 app entries, on **9,660 unique applications.**
- The variables are as follows:

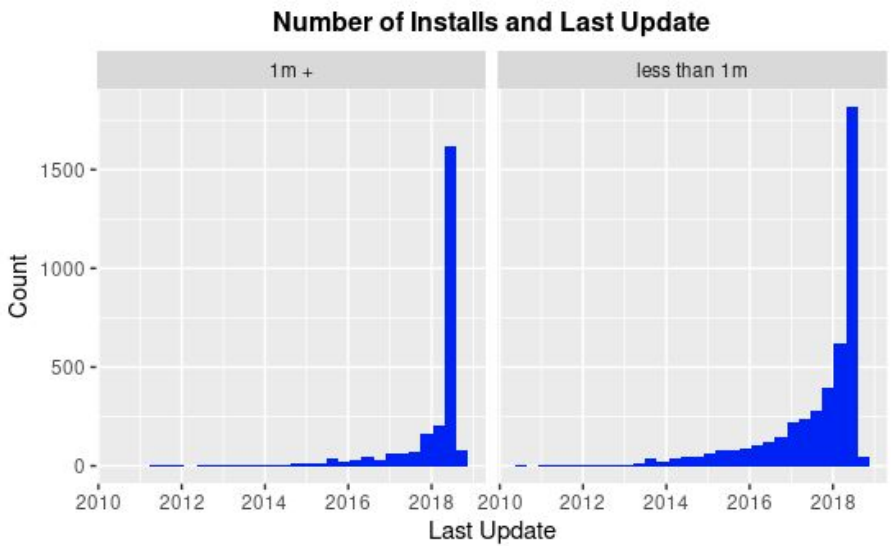
App Name	Category	Rating
Reviews	Installs	Size
Type	Price	Content Rating
Genres (More than one)	Last Updated	Current Ver
	Android Ver	

How we Prepared the Data



Key Insights from Exploratory Data Analysis





- Apps that have 1 million + installs tend to have higher ratings, particularly above 3
- Apps with 1 million + installs tend to have a greater number of reviews than apps that have less than 1 million installs.
- Apps that tend to have a higher rating tend to have a greater number of reviews too.

- A majority of apps with 1 million + installs tend to be the ones updated around early 2019.
- A great number of applications with less than 1 million installs had been updated in the past year.

Final Model

Logistic Regression Model

Variables of Interest

Interaction Terms

Full model

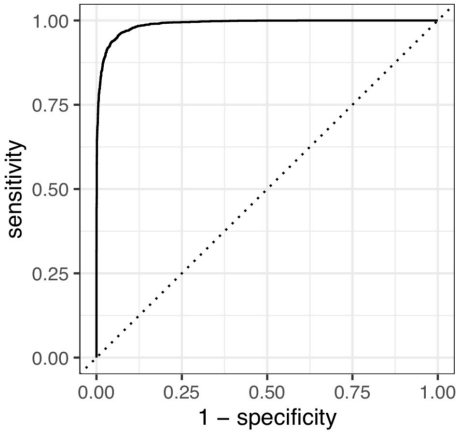
backward selection with BIC

adjusting for violated linearity assumption

detecting multicollinearity

Final Model

term	estimate	std.error	statistic	p.value
(Intercept)	-35.947	4.138	-8.688	0.000
Rating	10.159	1.920	5.292	0.000
log_reviews	2.201	0.074	29.653	0.000
TypePaid	-3.554	0.385	-9.228	0.000
Content_RatingEveryone	2.837	1.603	1.769	0.077
Content_RatingEveryone 10+	1.657	1.626	1.019	0.308
Content_RatingMature 17+	1.732	1.623	1.067	0.286
Content_RatingTeen	1.674	1.609	1.040	0.298
log_daysfrom	-0.347	0.096	-3.622	0.000
I(Rating^2)	-1.518	0.243	-6.240	0.000



AUC = 0.9889

Misclassification rate = 0.05 (threshold = 0.5)

Key Findings

The Model

$$\log(odds\ of + \widehat{1million\ downloads}) = -35.947 + 10.159 \times Rating - 1.518 \times Rating^2 + 2.201 \times \log(Number\ of\ Reviews)$$
$$-3.554 \times Type(Paid = 1, Free = 0) + 2.837 \times Everyone + 1.657 \times Age\ over\ 10 + 1.732 \times Mature(17+) + 1.674 \times Teen$$
$$-0.347 \times \log(Days\ from\ last\ update)$$

A set of significant predictors appropriate for **forecasting** whether an app gets 1 million + downloads

Suggests a **strategy** an app can undertake that correlates with a greater odd of getting more than a million downloads: **increase number of reviews and rating, have the application as free to install, and update the app frequently.**

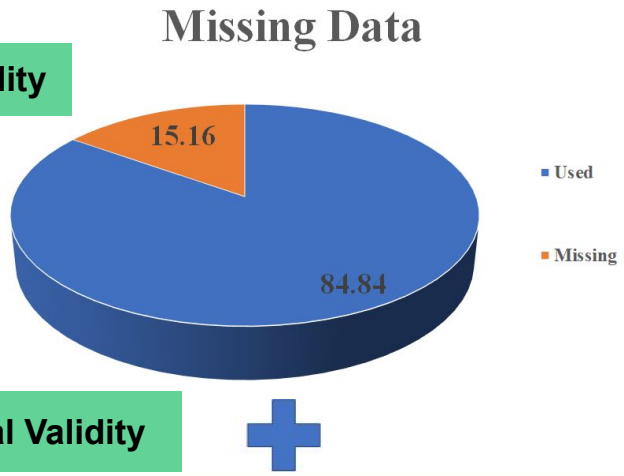
The Threshold

- The best decision making threshold we chose from the ROC curve is 0.5.
- This threshold gives us a sensitivity and specificity beyond **93.7%.**
- We chose a threshold that is both high in sensitivity and specificity because we find it important that both our Type I and Type II errors are low for our purpose of prediction.
- Suggests that **an app** with certain traits **could be tested** under this threshold to judge whether they are correlated with getting more than a million downloads.

Scope of Improvement

Dataset

Reliability



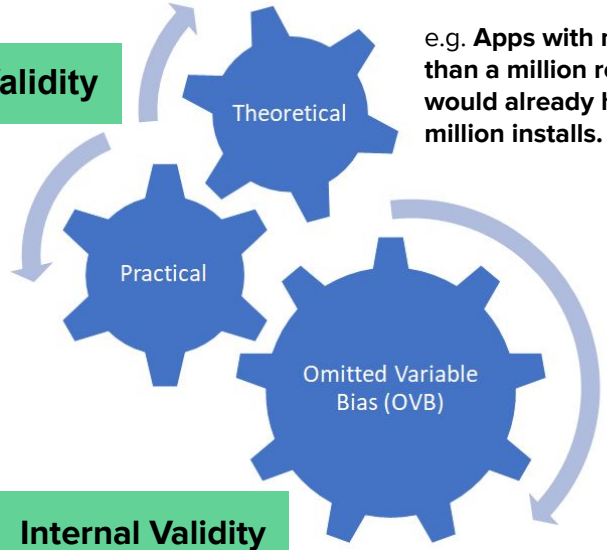
External Validity

Not a single popular social network application included

- Facebook, Twitter, Instagram
- Could suggest lack of representativeness

Variables

Construct Validity



e.g. **Apps with more than a million reviews would already have + 1 million installs.**

e.g. **If a user installed an app but deleted it later, it would still count towards the number of installs.**

Internal Validity

e.g. **Advertisement on the app, availability of the app other than New York, USA.**