The University of Chicago Booth School of Business

BUSN 41201 - Big Data - Final Project

PROJECT TITLE

26 May 2024

Yi Cao, Shri Lekkala, Ningxin Zhang

Contents

Note: The full the code used in all the questions can be found in the appendix.

1. Executive Summary

[REDO AFTER WE COMPLETE THE REPORT]

In this report, we present a comprehensive analysis of the "Google Play Store dataset" to gain insights into the characteristics and success factors of mobile applications. By examining various aspects related to app details, including categories, ratings, reviews, sizes, installations, and pricing, we aim to identify patterns and trends that contribute to an app's success on the Google Play Store.

We begin by exploring the general statistics of apps, focusing on the distribution of app categories, ratings, and reviews. This provides a foundational understanding of the data and highlights key areas of interest. Next, we delve into specific analyses to understand the relationship between app size, installs, and pricing, exploring how these factors influence an app's popularity and user engagement.

Our study also includes a sentiment analysis of user reviews, examining the polarity and subjectivity of feedback to understand how user sentiments correlate with app ratings and success. Additionally, we develop predictive models to forecast app ratings based on various features, and we investigate potential causal relationships between app characteristics and their performance metrics.

By leveraging data visualization, feature engineering, and predictive modeling techniques, we aim to provide actionable insights for potential app developers. These insights can help optimize app features, improve user satisfaction, and ultimately enhance the app's visibility and success on the Google Play Store.

2. Introduction

[WE CAN CHANGE THE QUESTIONS, THESE ARE JUST EXAMPLES]

In this paper, we aim to analyze the Google Play Store dataset to gain a comprehensive understanding of the factors that contribute to the success of mobile applications. The dataset includes details of apps such as categories, ratings, reviews, sizes, installations, and pricing, as well as user reviews with sentiment analysis. Our objective is to uncover patterns and trends that can help app developers optimize their offerings and improve user satisfaction.

The Google Play Store dataset, available on Kaggle, consists of two files: `googleplaystore.csv`, which contains detailed information about the apps, and `googleplaystore_user_reviews.csv`, which includes user reviews and sentiment data.

Our analysis will focus on the following research questions:

- **What factors affect the number of installs an app receives?** Specifically, what is the relationship between app size, type (free or paid), price, and the number of installs?
- **What are the key features that influence an app's rating?** How do factors like category, price, and number of reviews contribute to the overall rating of an app?
- **How does user sentiment in reviews correlate with app ratings?**
  Can sentiment analysis of user reviews provide additional insights into user satisfaction and app performance?

We will begin by loading and cleaning the dataset, followed by a thorough exploratory data analysis to uncover initial insights. Subsequently, we will perform detailed analyses to address our research questions, culminating in the development of predictive models and the identification of causal relationships. We will end by making concluding remarks from our research.

3. Dataset

a) Understanding the data

For `googleplaystore.csv` there are the following columns:

- App: Application Name
- Category: Category Type (e.g. Family, Game, Art)
- Rating: User rating review
- Reviews: Number of reviews
- Size: Download size of application
- Installs: Number of user downloads
- Type: Paid or Free
- Price: Price of App
- Content.Rating: Age group that app is targeted at (E.g. Everyone, Teen, Child)
- Genres: Other categories the app belongs to, other than the main category
- Last.Updated: Date when app was last updated
- Current.Ver: Current app version available
- Android.Ver: Minimum required Android version for app

There are a total of 10841 rows (applications).

For `googleplaystore_user_reviews.csv` there are the following columns:

- App: Application Name
- Translated_Review: User review, translated to English
- Sentiment: Positive / Negative / Neutral (Preprocessed)
- Sentiment_Polarity: Sentiment polarity score (Preprocessed)
- Sentiment_Subjectivity: Sentiment subjectivity score (Preprocessed)

This dataset contains the first 100 'most relevant' review for each app, with some prepocessing already done to add the last 3 features.

There are a total of 64295 rows (reviews).

b) Data Cleaning

For the `googleplaystore` dataset, we first process the variables by converting columns to the appropriate datatype. For example Installs, Size, Reviews Price, and Android.Ver are converted to numerics,

Last.Updated is converted to date. Then we filter out apps with Type 0 or NA, and remove duplicated rows.

After this, we are left with 10356 rows.

With the `googleplaystore_user_reviews` dataset, the variables were already well structured, but we noticed there were many rows with "nan"s. After filtering these out, we were left with 37432 rows.

4. Exploratory Analysis

c)

d)

e)

f)

5. What factors affect the number of installs an app receives?

A. Introduction

B. Analysis

Model 1.

Model 2.

Model 3.

C. Conclusion

6. What are the key features that influence an app's rating?

A. Introduction

B. Analysis

Model 1.

Model 2.

Model 3.

C. Conclusion

7. How does user sentiment in reviews correlate with app ratings?

A. Introduction

B. Analysis

Model 1.

Model 2.

Model 3.

C. Conclusion

8. Conclusion

9. Appendix

```
########################################
# Setup
########################################

knitr::opts_chunk$set(
    echo = FALSE,
    fig.height = 4,
    fig.width = 6,
    warning = FALSE,
    cache = TRUE,
    digits = 3,
    width = 48
)

# Required Packages
library(tidyverse)
library(ggplot2)
library(dplyr)
library(corrplot)
########################################
# 3. a) Understanding the datasets
########################################
# Load the datasets
googleplaystore_raw <- read.csv("data/googleplaystore.csv")
googleplaystore_user_reviews_raw <- read.csv("data/googleplaystore_user_reviews.csv")

# Check the column names
colnames(googleplaystore_raw)
colnames(googleplaystore_user_reviews_raw)

# Check the dimensions
dim(googleplaystore_raw)
dim(googleplaystore_user_reviews_raw)
########################################
# 3. b) Data Cleaning
########################################

# Convert the variables to the appropriate data type
googleplaystore <- googleplaystore_raw |>
  mutate(
    # Transform Installs and size to numeric
    Installs = gsub("\\+", "", as.character(Installs)),
    Installs = as.numeric(gsub(",", "", Installs)),
    Size = gsub("M", "", Size),
    # Convert apps with size < 1MB to 0, and transform to numeric
    Size = ifelse(grepl("k", Size), 0, as.numeric(Size)),
    # Transform reviews to numeric
    Reviews = as.numeric(Reviews),
    # Change currency numeric
```

```
    Price = as.numeric(gsub("\\$", "", as.character(Price))),
    # Convert Last.Updated to date
    Last.Updated = mdy(Last.Updated),
    # Change version number to 1 decimal, and add NAs where appropriate
    Android.Ver = gsub("Varies with device", NA, Android.Ver),
    Android.Ver = as.numeric(substr(Android.Ver, start = 1, stop = 3)),
  ) |>
  # Remove apps with Type 0 or NA
  filter(Type %in% c("Free", "Paid")) |>
  # Remove duplicate rows
  distinct()

# Remove all rows with nans
googleplaystore_user_reviews <- googleplaystore_user_reviews_raw |>
  filter(Translated_Review != "nan")
```