Google Play Store Apps

Abstract

This project aims to make Exploratory Data Analysis (EDA) and prediction models on the Google Play Store Apps dataset. These EDA and models will help the developers to understand the type of application people is preferred. I worked on a dataset founded through the Kaggle website. I used python libraries such as NumPy, pandas, and Matplotlib.

Design

The data has been collected by using Python script in June 2021. By Applying EDA the following questions will be answered:

- Will the price affect number of installations?
- What is the most downloaded app?
- What is the most famous category of the app?

Data

The dataset contains over 2 million instances with 22 features for each. Features include numerical and categorical types, such as rating, type, developer, etc. By using a linear regression model I will be able to predict the missing values.

Algorithms

- Feature Engineering
  - Handle missing values in rating and minimum android columns.
  - Drop some non-useful columns such as Currency, Size.
  - Factorizing categorical features to int variables, such as Category and Free.

- Models
  - Classification model to classify App installs.
  - Regression model to predict the rate of future App.

Tools

- Numpy and Pandas to handle the dataset.
- Matplotlib and Seaborn to visualise the data.
- Sklearn to build the models