Mobile App Size, Rating & Installs Prediction

Group #14

Sanath Davis
10088657

Capstone Term 1

Final Project Presentation

13 Dec 2022

Bindia Biju
100886575

Capstone Term I
05/12/2022

Mobile App Size, Rating & Installs Prediction

Group **#14**

Bindya Biju
ID: 100886575

Sanath Davis
ID: 100884693

Abstract

The final project report document takes the reader through the final design of various experiments on the dataset and their results. It reiterates the motivation behind the project and delves into the details of the dataset. The existing projects based on similar datasets are also explored. A concluding note is also attained.

Keywords: google, play store, price, size, category, installs, ratings

1. Introduction

In this day and age of mobile apps, every small decision made by mobile app developers will affect the sales of an app. There are many crucial decisions to be made, like the price of an app or the size of an app, which would lead to maximum profits, better user ratings, and the maximum number of installs. These decisions might vary depending on the category of the app.

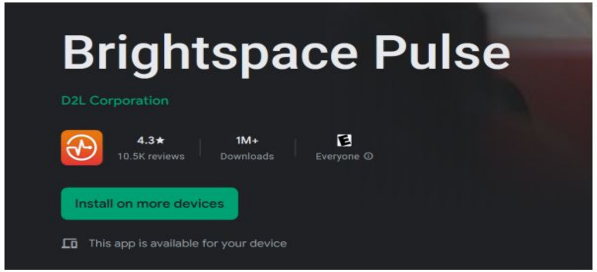


Figure 1. Details of an app in Google Play Store (Representative image)

After our machine learns this dataset, it would then be able to predict the ratings and number of installs an app would get given a particular size, category, and price. Also, we

1.

Motivation

Target Audience : Mobile App Developers and Companies

Our project would be very helpful to Mobile app developers to

- *Decide what their app size should be*
- *Predict the rating they will get*

Dataset Link:

Google Play Store Apps

https://www.kaggle.com/datasets/gauthamp10/google-playstore-apps

Problem Statements

- What should be the size of an app to get more installs?
- Predict the Rating of an app
- Predict the Number of Installs of an app

2.

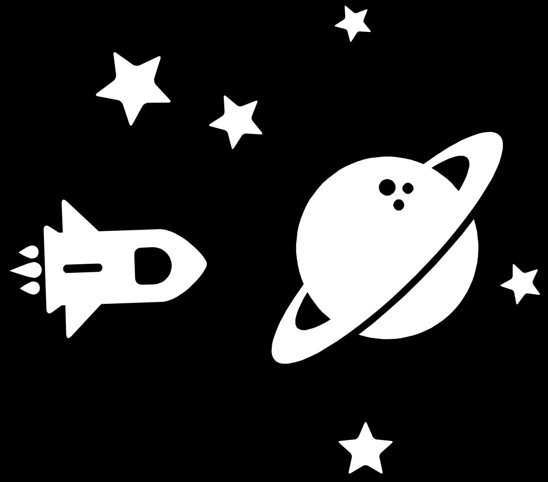
Related Works

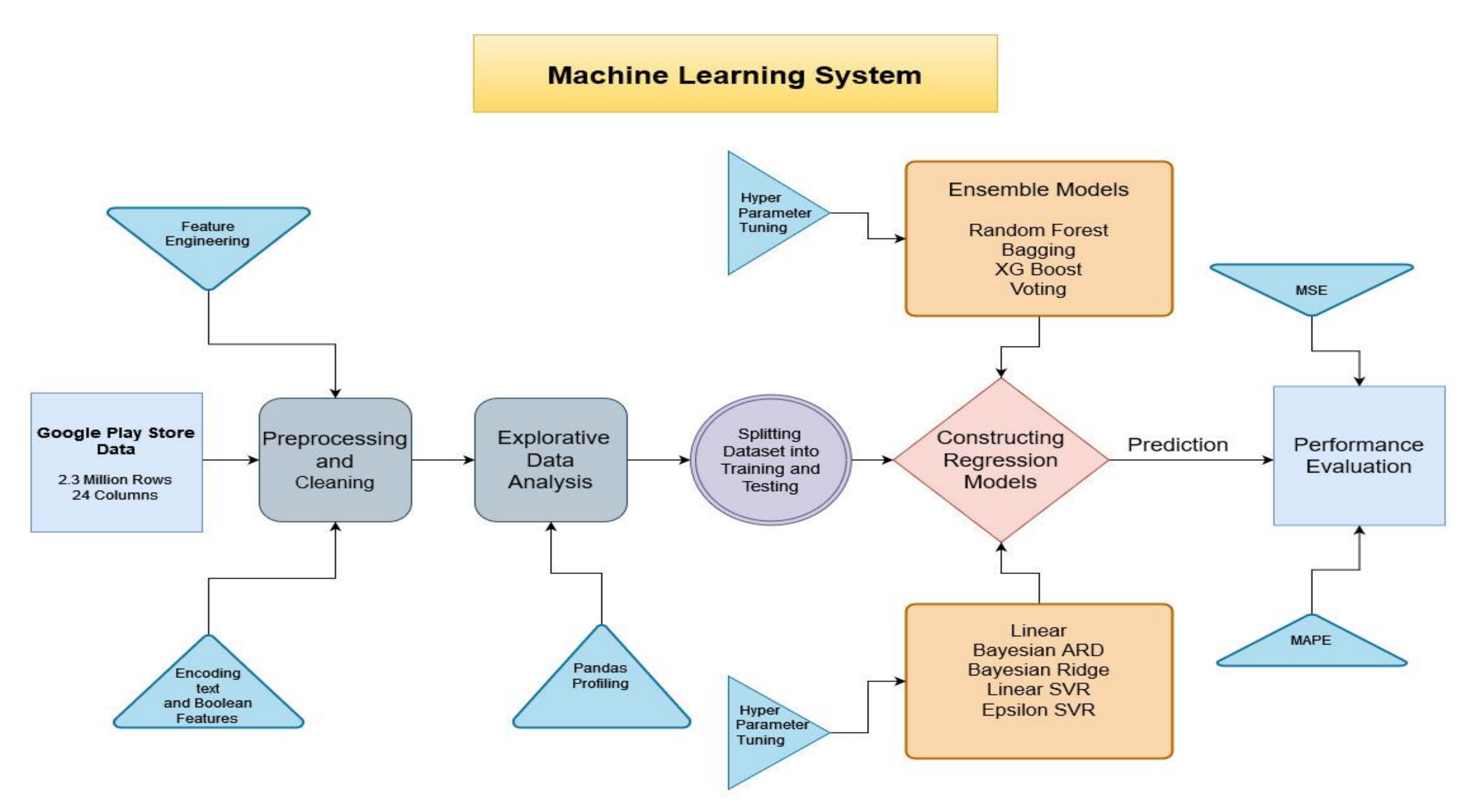
Why our work stands out

- Most of the related works referenced in our paper are doing various data analyses
- But most are not doing actual prediction
- Also, our project compares a large number of models including ensemble models and we do Hyperparameter tuning too
- We are also focusing on less common features like Size, Rating and Installs

OUR APPROACH

Ideas, methods, designs




Machine Learning System
Feature Engineering
Google Play Store Data
2.3 Million Rows
24 Columns
Preprocessing and Cleaning
Explorative Data Analysis
Encoding text and Boolean Features
Pandas Profiling
Splitting Dataset into Training and Testing
Hyper Parameter Tuning
Ensemble Models
Random Forest
Bagging
XG Boost
Voting
Constructing Regression Models
Prediction
Linear
Bayesian ARD
Bayesian Ridge
Linear SVR
Epsilon SVR
Hyper Parameter Tuning
MSE
MAPE
Performance Evaluation

Exploratory Data Analysis

Box whisker plots

Density plots

Correlation

Pandas Profile Report

Scatter Plots

Histograms

Data pre processing

Cleaning

Remove unwanted columns

Remove unnecessary symbols

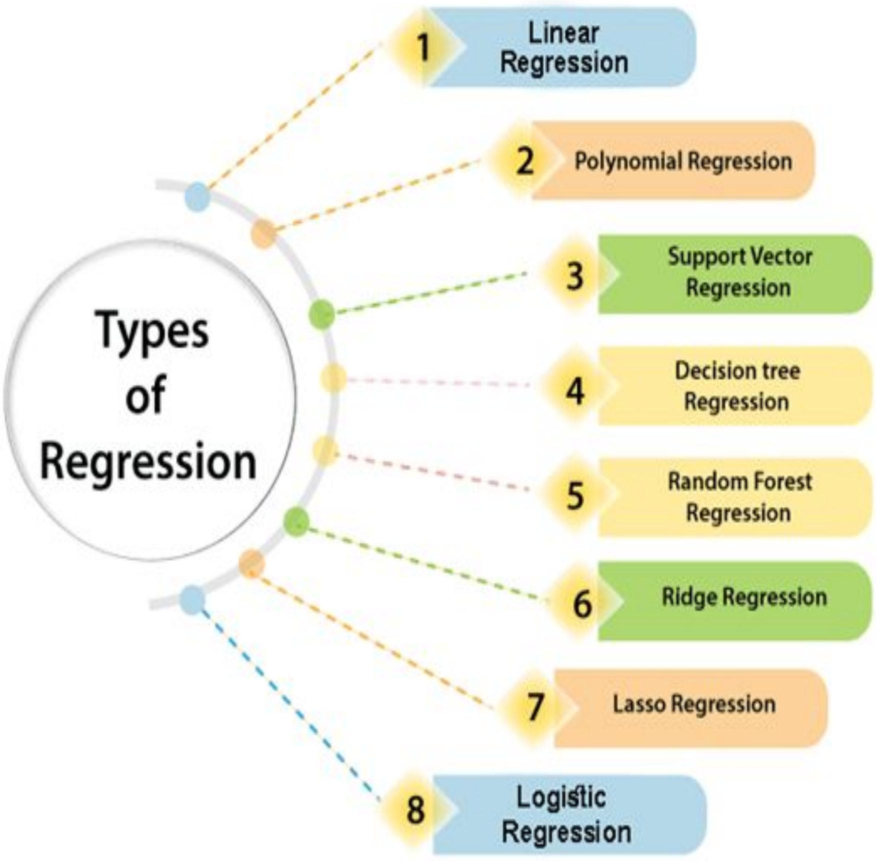
Feature Engineering

Create new relevant features from existing features

Encoding

Convert strings and booleans to numeric data type for conduction regression

Modelling



Types of Regression
1 Linear Regression
2 Polynomial Regression
3 Support Vector Regression
4 Decision tree Regression
5 Random Forest Regression
6 Ridge Regression
7 Lasso Regression
8 Logistic Regression

Regression Models used

➢ Predicting App Ratings

Regression
Linear (Baseline)
Random Forest
Bayesian ARD
Bayesian ridge
Linear SVR
Epsilon SVR

Ensemble Model
Random Forest
Bagging Regressor
XG Boost Regressor
Voting Regressor

Conclusion and Insights

Category	Minimum Size	Maximum Size	Best Size Range to get Maximum Installs
Adventure	0.045 MB	1100 MB	800-900 MB
Maps & Navigation	0.016 MB	382 MB	150-200 MB
Role Playing	0.043 MB	1500 MB	1000-1200 MB

Figure 6. Best Size range in different categories

Random Forest

In our observation, Hyper Parameter Tuned Random Forest Model gave the best predictions

➢ Predicting App Ratings

Regression	MSE	MAPE
Linear (Baseline)	4.221	4354160987210370.0
Random Forest	0.4	0.07
Bayesian ARD	2.058	4369784657954465.5
Bayesian ridge	2.054	4354193448350276.5
Linear SVR	2.508	3135247702318329.5
Epsilon SVR	Taking too much time to fit	Taking too much time to fit

Figure 7. Results from multiple regression models

Ensemble Model	MSE	MAPE
Random Forest	0.4	0.07
Bagging Regressor	2.08	5121046823101658
XG Boost Regressor	0.45	10671825510858
Voting Regressor	0.45	10448593663074

Figure 8. Results from different ensemble regression models

Tuned Model	Best Parameters	MSE Improvement	MAPE Improvement
Random Forest (Randomised Search)	{'n_estimators': 200, 'min_samples_split': 5, 'min_samples_leaf': 4, 'max_features': 'auto', 'max_depth': 10, 'bootstrap': True}	3.2 %	4.08%
Support Vector Regressor (Grid Search)	{'C': 10.0, 'gamma': 0.01}	50%	70%
XG Boost Regressor (Randomised Search)	{'subsample': 0.8999999999999999, 'n_estimators': 1000, 'max_depth': 3, 'learning_rate': 0.01, 'colsample_bytree': 0.8999999999999999, 'colsample_bylevel': 0.7999999999999999}	.01 %	175%

Figure 9. Improvements from Hyper-parameter Tuning

Key Findings

1

All models performed better than the baseline

2

Random forest gave the least error rates while predicting app ratings

3

All ensemble methods we tried gave low MSE rates

4

Hyperparameter tuning improved prediction performances.

5

MSE and MAPE are useful in comparing models. Some models are not suitable for the dataset

6

There might be other features which influence the app ratings and number of installs

Thanks!

Any questions?

You can find us at

- sanath.davis@dcmail.ca
- bindia.biju@dcmail.ca

PPT Template Credits

Special thanks to all the people who made and released these awesome resources for free:

- Presentation template by SlidesCarnival
- Photographs by Unsplash