CSE 4990/6990 – BIG DATA AND DATA SCIENCE

PREDICTING GOOGLE APP RATING

Team: 5DBminds
Presenter: Naila Bushra

PROJECT OVERVIEW

“Predicting PlayStore Rating for Apps”

Using PlayStores data of a large number of existing applications

To predict ratings for future apps based on the attributes of that application

DATA COLLECTION

The data of this project has been collected from the author of the github repository GooglePlayAppsCrawler

GitHub Link: https://github.com/MarcelloLins/GooglePlayAppsCrawler.git

DATA COLLECTION


```
"AppSize":-1.0,
"Category":"SOCIAL",
"ContentRating":"Rated 12+",
"CoverImgUrl":"https://lh3.googleusercontent.com/ZZPdzvlpK9r_Df9C3M7j1rNRi7hhHRvPhlklJ3lfi5jk86Jd1s0Y5wcQ1QgbVaAP5Q=
"CurrentVersion":"Varies with device",
"Description":"Keeping up with friends is faster than ever.• See what friends are up to• Share updates, photos and \
"Developer":"Facebook",
"DeveloperEmail":"android-support@fb.com",
"DeveloperNormalizedDomain":null,
"DeveloperPrivacyPolicy":"https://www.facebook.com/about/privacy/\u0026sa=D\u0026usg=AFQjCNGsgQ5qAO5ohRTZIcLRwgVSGnk
"DeveloperURL":"/store/apps/developer?id=Facebook",
"DeveloperWebsite":"facebook.com",
"HaveInAppPurchases":false,
"Instalations":"1,000,000,000 - 5,000,000,000",
"IsFree":true,
"IsTopDeveloper":true,
"LastUpdateDate":{ ⊞ },
"MinimumOSVersion":"Varies with device",
"Name":"Facebook",
"PhysicalAddress":"",
"Price":0.0,
"PublicationDate":{ ⊞ },
"ReferenceDate":{ ⊞ },
"RelatedUrls":[ ⊞ ],
"Reviewers":-1.0,
"Reviews":[ ⊞ ],
"ReviewsStatus":"Visited",
"Score":{ ⊟
    "Count":3.0383292e+07,
    "FiveStars":0.0,
    "FourStars":0.0,
    "OneStars":0.0,
    "ThreeStars":0.0,
    "Total":3.99822998046875e+14,
```

DATA HANDLING

The data itself is in a json format, to use it in our project we have

Split it into smaller data chunks

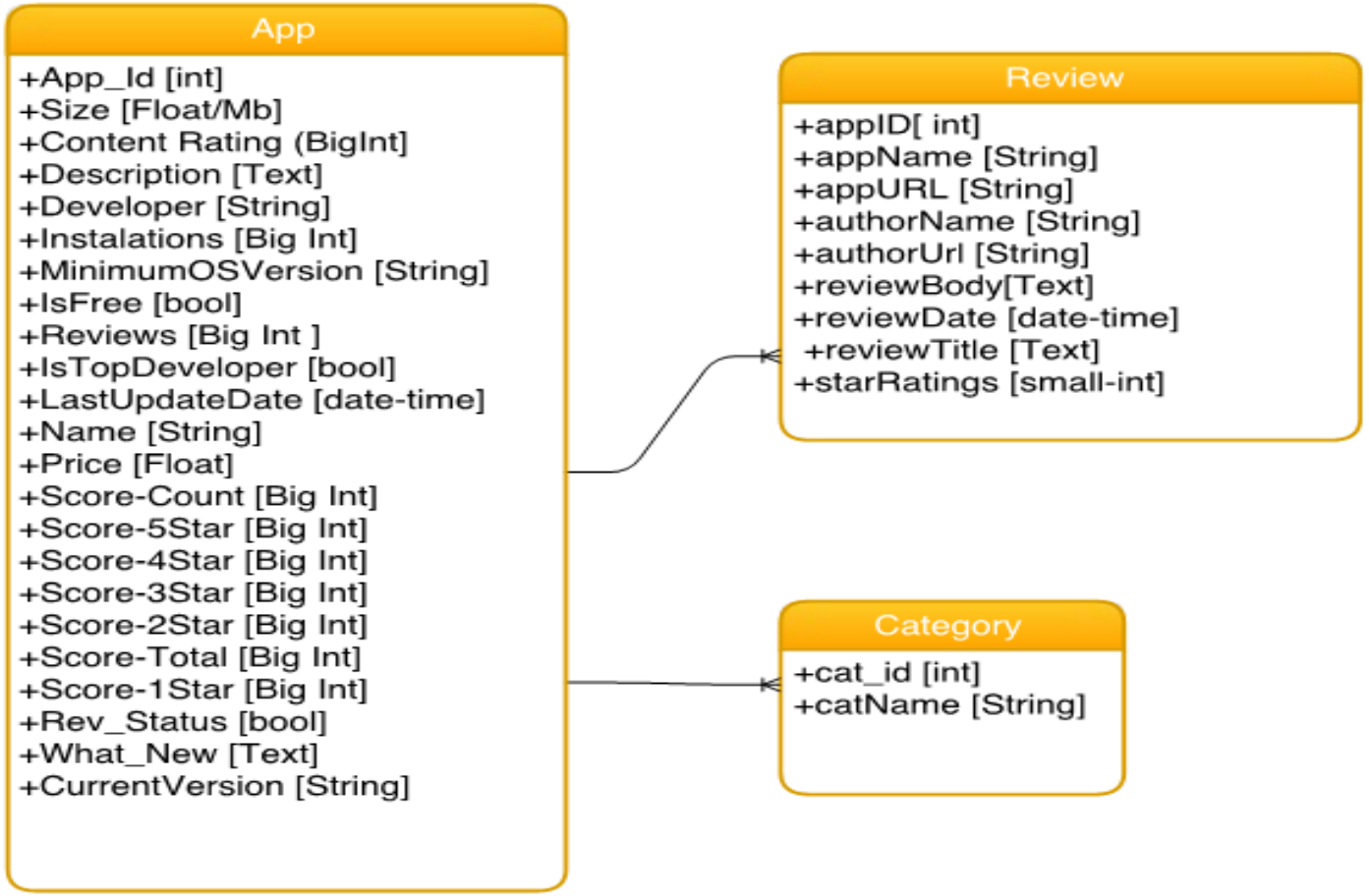
Loaded into python data structure dictionary

Assigned it to pandas dataframe using the attributes as columns

DATA PREPROCESSING

AppSize
Category
ContentRating
Developer
Description
HaveInAppPurch
ases
Instalations
IsFree
IsTopDeveloper
LastUpdateDate
Name
Price
PublicationDate
Reviewers
Reviews
ReviewsStatus
Score

ER DIAGRAM OF GOOGLE APP RATING DATA



ISSUES AND RISKS

Some of the attributes values are not numerical

We are currently dealing with the data of nearly 11,000 apps so our biggest concern is the computational time!

NEXT PHASE

Assign numerical values to the Application features

Standardize the data using standard score

Construct feature vectors

Apply supervised learning algorithm to build a prediction model using the training data

Split the data into 70/30 ratio for train and test data.

To avoid overfitting we will randomize the train and test data using approaches like k fold cross validation/ random sub-sampling validation etc.

THANK YOU