PREDICTIVE ANALYTICS AND VISUALIZATION FOR iOS APPSTORE

By

Rutuja Shinde

Vinaya Chinti

Shivani Ghatge

Laxmi Prasanna Gundabolu

Course Project Professor's Name

Vadim Sokolov

Course Name

Applied Predictive Analytics

Course Number and Section

SYST/OR 568

Date

4/10/20

GEORGE
MASON
UNIVERSITY

OBJECTIVE

This project focuses on analyzing various features that contribute to the success of the applications in the app store for iOS devices. We intend to build a regression model for prediction of the user ratings for applications and a classification model to predict the genre of the application.

SELECTED DATASET

This dataset was extracted from the **iTunes Search API** at the **Apple Inc website**. We have accessed this data from Kaggle. It comprises of **7197 rows and 16 columns** of Apple iOS mobile application details. The following are the contents in each of these files :

appleStore.csv

- id" : App ID
- "track_name": App Name
- "size_bytes": Memory Size of App (in Bytes)
- "currency": Currency Type
- "price": Price of App
- "rating_count_tot": User Rating counts (for all version)
- "rating_count_ver": User Rating counts (for current version)
- "user_rating" : Average User Rating value (for all version)
- "user_rating_ver": Average User Rating value (for current version)
- "ver" : Latest version code
- "cont_rating": Age group eligible for access of App
- "prime_genre": Primary Genre of App
- "sup_devices.num": Number of supported devices
- "ipadSc_urls.num": Number of screenshots showed for display
- "lang.num": Number of supported languages
- "vpp_lic": Vpp Device Based Licensing Enabled

EXPLORATORY ANALYSIS

The exploratory analysis will be used to answer the following questions
1) Which application is most popular based on the total ratings of the application?
2) Do customers give a good rating to the apps if the size of the applications is more?
3) Do customers prefer paid applications or free applications?
4) Which genre of applications is mostly used by the users?
5) Does the price of the applications depend on the genre of the application?
6) Does the price of the applications increase with the size of the application?

MODELS

The target variable will be “user_rating” for prediction and for classification the target variable will be “prime_genre”. Following algorithms will be used for prediction and classification.

- Support Vector Machine
- Regression Model
- Random Forest

The data does not have any NA’s so we do not have to delete any rows from the database. The non-English names of the applications will be removed from the data frame. For exploratory analysis purposes, we will add a column to the data frame which will have two values “paid” or “free”. If the price column has a value greater than 0 then the value of the new column will be “paid” otherwise the value of the column will be “free”. The same can be done for the column cont_rating, this column specifies the age group of the users who can use the application. We will have a new column whose value will be “teen” if the age is between 4 to 16, and “adult” if it is 17+. Predictive analytics can be applied to the data to predict the rating of the application which can be achieved by using a regression algorithm. The classification algorithm can be used to classify the application into different genres. This will be a multi-class classification problem as we have multiple genres for the classification. For multi-class classification, we will use algorithms like random forest and support vector machines. A comparative analysis will be done based on the accuracy of both the algorithms. The data will be split into a 70-30 ratio of which 70% will be used for training the model and the predictions will be done on the remaining 30% of the data. Cross-validation will be used as well as we will calculate the confidence interval and use various statistical tests to evaluate the accuracy of the models built. For the regression model the “user_rating” column will be the target variable and for the classification model “prime_genre” will be the target variable.