PROJECT PROGRESS REPORT

CSE-4990/6990 - Big Data and Data Science

Predicting Google App Rating

Team 3: 5DBminds

Date: November 5, 2015

Ayush Raj Aryal
Lucas Andrade Ribeiro
Naila Bushra
Naresh Adhikari

Previous Status:

Previously, we had our play store application data in JSON format which was converted into csv format along with the removal of unnecessary data. We were planning to look at assigning numerical values to the application features and then apply various supervised learning algorithms to build a model for prediction.

Current Updates:

Data Transformation:

As all of the features of our feature set are not numerical, we have performed following transformations:

1. **Features with numerical value**

   Features such as "AppSize" and "Price" already have numerical values so these are used directly in the feature vector. "LastUpdateDate" and "PublicationDate" are used to calculate the number of months since the corresponding application was updated and uploaded respectively. The feature "Installations" has a numerical range as feature value which has been splitted into minimum value of the range and maximum value of the range.

2. **Features with textual value**

   Features like "HaveInAppPurchases", "IsFree", "IsTopDeveloper", "ContentRating" were converted to 0 or 1 depending upon their string value True and False.

   The features such as "Name", "Developer", "Description" contain texts so they are vectorized using sk-learn's word count vectorizer which gives a 1D vector for each text descriptor. However, to make our conversion effective NLTK library is used which helped us to do the vectorization based on the stem/root of words (For example go for going/went/gone)

   In case of "Category" there are 42 unique categories identified in the whole dataset. To convert it into numerical values, incremental integer numbers were assigned to them for example 1 for the GAME, 2 for the MUSIC and so on.

Present Issues:

In case of vectorization the number of columns in the vectorized matrix is very high which might result in memory error or very high processing time. Also, the Description columns contains many Unicode characters for languages other than English which need some sort of transformations to be included for feature vectors.

In case of Category, our methodology to assign incremental numerical values will result in assigning different weights to different categories which can bias our prediction results. A better method to assign weights needs to be identified.

Task Division:

Currently we have divided our project tasks within ourselves in the following manner:

a. Text Vectorization : Naresh Adhikari
b. Linear Regression : Naila Bushra
c. Naïve Bayes: Ayush Raj Aryal
d. Clustering : Lucas Andrae Ribeiro

Next Phase

After we have fully processed our data in a standard format, we are planning to implement the following steps:

- Supervised learning approaches (Linear regression, Naïve Bayes, K Nearest Neighbor) to build prediction models using our training data
- To avoid overfitting we will randomize the train and test data using approaches like k fold cross validation/ random sub-sampling validation etc.