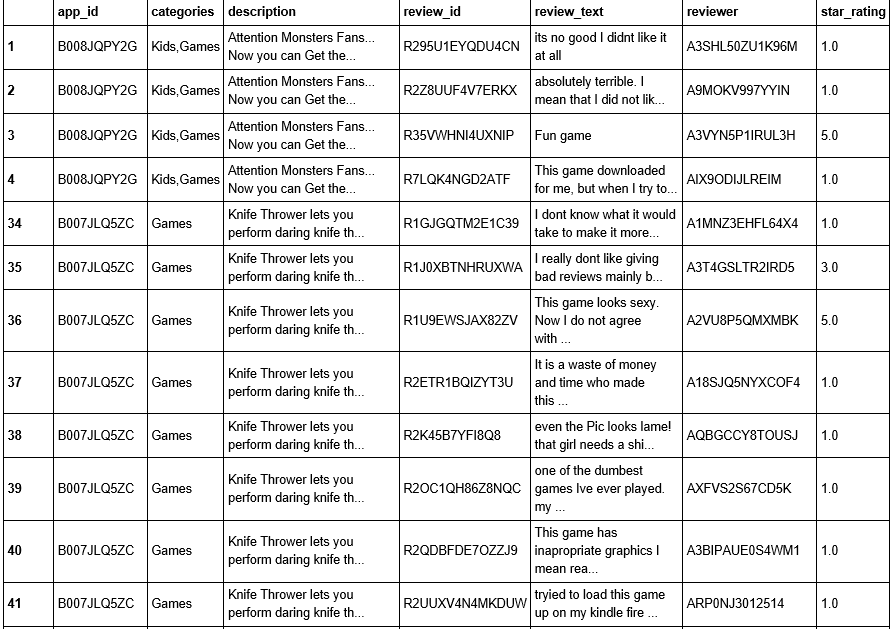
CS 591 Progress Report 1

Leshen Sui, Zheming Sun

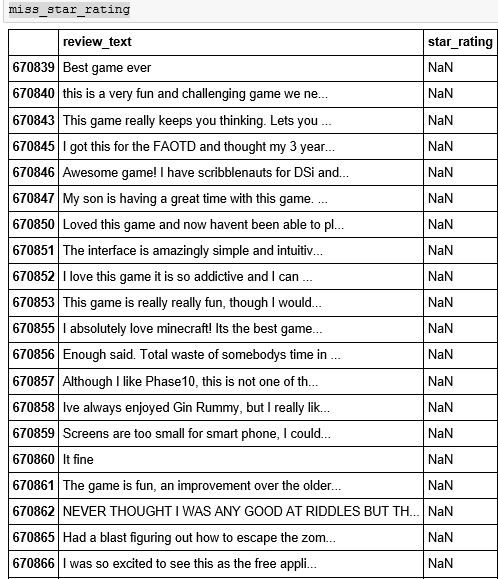
11/11/2016

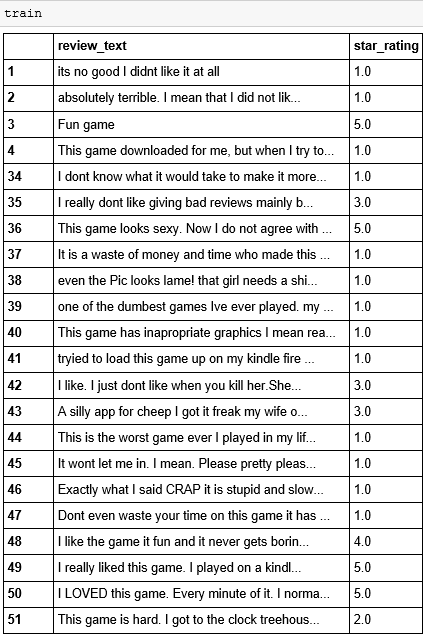
What we have done so far was data pre-processing and dealing with missing values with classification techniques. As I mentioned in my proposal, our row data set contains information about Amazon paid apps and we want to extract useful features and only focus on game apps. After doing feature selection and category filtering our data set is like:



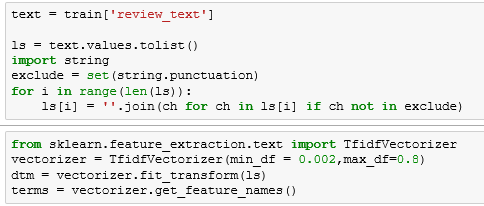
Notice that now categories are all about games.

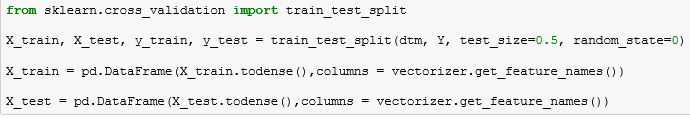
Then we focus on missing values. It turns out that missing values only show up in review\_text and star\_rating columns, so we decided to use classification techniques to predict missing values by reviews. In order to achieve this, I delete all records that don’t have a review(nan in review\_text column), and then separation our data set two part based on whether star\_rating is missing:



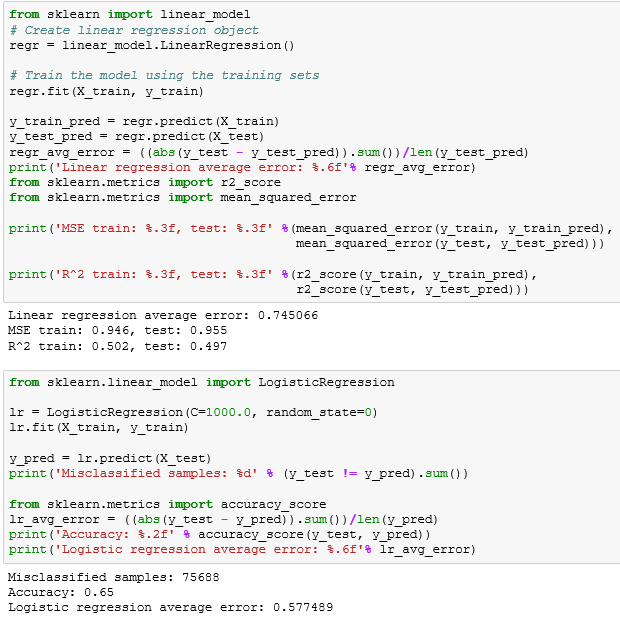
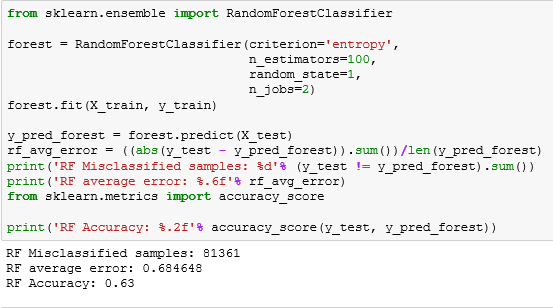
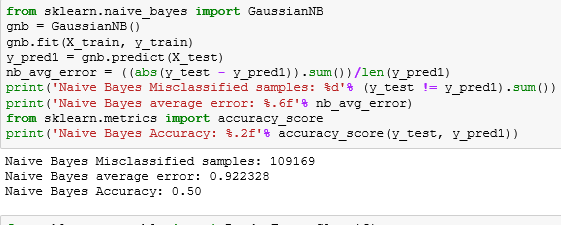


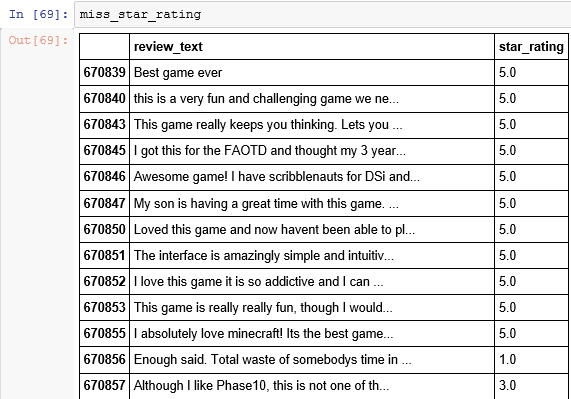
What I did next is to vectorize review\_text columns, and then split the “train” data frame into training set and test set to construct a classification model. After fitting the model we do evaluation based on the average difference between the true star\_rating and predicted staring\_rating. The intuition is we want our predicted star\_rating is close to the true star\_rating; for example, predicting a 5 star to a review when the true rating is 4 star is acceptable, but we want to avoid that predicting 1 star rating as 5 star. Therefore avg(|ytrue - ypredicted|) should be as small as possible.





We tried four different classification model in total, and logistic regression gives the best result:

 The average error of Logistic regression model is 0.58, which means that if I predict a star\_rating as x star based on logistic regression model, the true value should fall in the range (x - 0.58, x + 0.58) on expectation, which is acceptable result. Also, the accuracy score is 0.65, which means 65% of the test data are prediction exactly correct. Therefore, we are going to use logistic regression to predict our missing values.



Looks like the result of prediction make sense!

Until now the data pre-processing has been completed. Our next step will be clustering apps based on descriptions, and label each cluster by some key words.