DSCI553 Recommendation System Competition Project

My implementation of the recommendation system primarily uses Singular-Value-Decomposition (SVD), XGBoost, Up-sampling, User profile (Friends).

The hybrid approaches I used include Mixed Hybrid and Feature Augmentation.

Training process:

Feature Augmentation is for the SVD model. Since we only have limited amount of training data, to better train the SVD model, I up-sampled the training set by randomly adding user-business-avg star pairs to current training set. Adding 300,000 pairs will yield a better result without overfitting the model.

Then I trained the rest of aforementioned models separately and save them to files.

Prediction process:

In the prediction process, the primary model will be XGBoost, which is good enough for most of the testing pairs, which is seen/known user and seen/known business. But there are several cases that XGBoost cannot make a good prediction on:

- Case 1: Unseen business with seen user
    - Predict the ratings through the friends of the known user. If this user has enough friends, we can predict the rating by taking the average ratings of all the friends. This approach assumes that the user will have similar taste with his/her friends (Otherwise they might not be friends). If this user did not have any friends on records, we can simply predict the rating by using this user's average rating over all businesses.
- Case 2: Unseen user with seen business
    - Predict the ratings through the friends of the user. Because the absence of the ratings of a user does not imply the absence of his/her friends. We can still use similar way to predict rating as in case 1. If the user does not exist in friends model, then we can use the average rating of this business as the predicted rating because the average rating of a business, although might not reflect this unknown user's rating accurately, is still representative of the business's level and therefore generate roughly accurate prediction.
- Case 3: Both user and business are unseen
    - This means we do not have any information on this user nor this business. This is a completely new pair. I predicted this by using weighted hybrid of average stars of all users and average stars of all business. Both will have a 0.5 weight.

The SVD and XGBoost will both predict on the same dataset and generate two prediction result, then I did a weighted sum of these two prediction results. SVD result will have a weight of 0.15 and XGBoost result will have a weight of 0.85. This way, I combined different recommenders together to generate a result jointly (Mixed Hybrid).