ST 810 : Final Project - Yelp Rating Prediction

Yingzi Xu

1 Problem Description

This project is about personalized business recommendations for Yelp user. To be more specific, we are aiming to creat a model to predict the rating a user would assign to a business based on a detailed snapshot of Yelp data, which are over 10,000 businesses, 8,000 check-in sites, 40,000 users, and 200,000 reviews from the Phoenix, AZ metropolitan area.

2 Summary of Findings

Among all the six models shown in Table 1, blended Model did the best rating prediction in terms of Root Mean Square Error (RMSE). It gives a RMSE of 1.137 and improves 6.5% compared with the overall mean benchmark.

Model	Detail	RMSE	Improvement
Mean Model	Overall Mean	1.217	0
	User Mean	1.266	- 4.0%
	Business Mean	1.153	5.3%
	User-Business Mean	1.219	-0.1%
Business-Based Collaborative Filtering Model		1.275	-4.8%
Blended Model		1.137	6.5%

Table 1: Model RMSE Performance Comparison

3 Data Description

1) In the given data, we have 11,537 businesses, 8,282 checkins, 43,873 users, 229,907 reviews. I would only focus on the recview data set here, since other information like 'business name', 'business location', 'votes' , 'checkin number during each time period', seems has no significant effect on user rating from my point of view.

2) Review data set are then randomly split it into two subsets: 80% training (183925 records), 20% testing (45982 records).

4 Analysis

After pre-processing of the data, I create several models as following.

4.1 Mean Model

I first generated four different kinds of mean models to give a basic idea.

1) **Overall Mean Model**
Here, I assume all the users would give a same rating for all the businesses. This is a very strong assumption which could not be true in the real world, however, this kind of starting point is a benchmark for other models to compare with. The model is,

$$\hat{r_{ui1}} = \bar{r}, \tag{1}$$

and it leads to a RMSE of 1.217.

2) **User Mean Model**
I then tried the user-mean model. If I could find user u in the training data set, I would use the average rating this user has given to all the business as a predicted rating for that user. Otherwise, overall mean would be applied.

$$\hat{r_{ui2}} = \begin{cases} \bar{r_u} & \text{if user } u \text{ is in training set} \\ \bar{r} & \text{otherwise} \end{cases} \tag{2}$$

This gives us RMSE of 1.266. The assumption here is, for one specific user, he would give a same rating to all the businesses. However, if we have no information about that user, I would treat him as all the other users and they would give the same rating to all the businesses.

3) **Business Mean Model**
After applying the user-mean, a natural thought is to try business-mean. If we know information about business i in the training data, we assume business i would receive same rating from all the users, which is given as,

$$\hat{r_{ui3}} = \begin{cases} \bar{r_i} & \text{if business } i \text{ is in training set} \\ \bar{r} & \text{otherwise.} \end{cases} \tag{3}$$

This model yeilds a RMSE of 1.153, which is much better than overall mean and user mean. It is reasonable since in the training data, we have about 20 average reviews for each business but only 6 average number of reviews for each user. The more information, the better result.

4) **User - Business Mean Model**
In this model, both user and business information would be utilized.

$$\hat{r_{ui4}} = \begin{cases} \bar{r} + u_u + b_i & \text{if user } u, \text{ business } i \text{ both in training set} \\ \bar{r_i} & \text{else if business } i \text{ is in training set} \\ \bar{r} & \text{otherwise} \end{cases} \tag{4}$$

It gives me a RMSE of 1.219, which performs worse than a simple business mean model.

4.2 Business-Based Collaborative Filtering Model

I only use business-based filerting rather than user-based filtering here since from the above four mean models we have had an idea that business mean would give a more accurate prediction in regarding of rating. In this model, I assume that for user u, the rating for item i is the weighted average of the same user's ratings on similar items k. which is formulized as,

$$\hat{r_{ui5}} = \begin{cases} \bar{r} + \frac{\sum_{k \in N_u(i)} w_{ik}(r_{uk} - \bar{r_k})}{\sum_{k \in N_u(i)} w_{ik}} & \text{if user } u, \text{ business } i \text{ both in training set} \\ \bar{r_i} & \text{if business } i \text{ is in training set} \\ \bar{r} & \text{otherwise} \end{cases} \tag{5}$$

In order to find similar items, I use Jaccard index as the similarity measure. A bigger Jaccard index between two businesses means more similar than others. Basicly, it is the size of interaction between two businesses divided by the size of union if these two businesses, given as,

$$J_{ik} = \frac{|I \cap K|}{|I \cup K|} \tag{6}$$

and then the weight I chose here is the same as Jaccard index. If two business are more similar, more weight would be put on.

$$w_{ik} = J_{ik} \tag{7}$$

However, one thing we do need to be careful is that Jaccard index would sometimes give an opposite similarity. In order to conquer this, I would calculate Jaccard index on rounding data rather than the raw data. Ratings of 3,4,5 are considered as '1', and ratings of 1,2 are treated as unrated. After doing the rounding, similarity measure is more accurate. The model finally get a RMSE of 1.275.

Sadly, it's worse than any mean model. This is first because of the high sparsity of user-business matrix and secondly, since 30% of the traing data are having ratings of 1,2, I lose some information when the rounding procedure is conducted.

4.3 Blended Model

Finally, I blended all the models above together and give weight as how well they perform on RMSE.

$$\hat{r_{ui6}} = \frac{\sum_k (w_k \hat{r_{uik}})}{\sum_k w_k}, \text{ where } k = 1, 2, 3, 4, 5 \tag{8}$$

$$w_k = \frac{1}{RMSE^3} \tag{9}$$

In the end, RMSE of 1.137 is achieved, which is my best model.