For this assignment, I extracted features from several json files. Due to word limit, I will just mention the innovative features I used to predict ratings apart from the most common features like user’s average rating, business review count, business’s average rating etc.

1. A feature I added is business\_rivalry score, I first grouped the businesses by the postal codes. After this, I grouped the businesses based on zipcode and accordingly set the rivalry score of the business as in if a business is in a famous place, its rivalry score is high.
2. 5. The average Yelping age of a user as in users who have been veteran Yelpers are given more weightage.
3. Number of days the business remains open in a week (from business.json)

4. Value of the “RestaurantsPriceRange2” value (from business.json)

5. Number of days the business remains open during weekends (from business.json)

6. “True” values of attribute “attributes” (from business.json). Here I extract the available features for a business by checking if the value of the attribute is True/False. I also considered the number of categories of a business.

As I had so many features, I tested and discovered that ensemble models performed the best. So I used HistGradientBoostingRegressor model and below are the answers:

HistGradientBoostingRegressor has inherent support for np.nan values so imputation isn’t mandatory. Also, HistGradientBoostingRegressor is much faster than other Gradient Boosting Regressor models for huge datasets.

For best results, I tuned the hyperparameters like max\_iter, max\_depth with values [100,5],[100,7],[200,8],[200,9],[400,9] to get the best results. Turns out the model performs best for max\_iter = 200 and max\_depth = 9.

Error distribution on validation dataset :

>=0 and <1: 102424

>=1 and <2: 32573

>=2 and <3: 6241

>=3 and <4: 806

>=4: 0

RMSE :

0.9789481969664

**Execution time :**

Duration : 378.4022331237793 seconds

1. User’s average rating (from user.json)
2. User’s review count (from user.json)
3. Value of “useful” attribute (from user.json)
4. Value of “funny” attribute (from user.json)
5. Value of “cool” attribute (from user.json)
6. Average rating of a business\_rating (from business.json)
7. Number of reviews received by the business (from business.json)
8. Value of “is\_open” attribute of business (from business.json)
9. Value of “NoiseLevel” attribute of business (from business.json)
10. Value of “WiFi” attribute of business (from business.json)
11. Number of times the business has been checked in (from checkins.json)
12. Number of photos of a business (from photo.json)
13. Number of likes received by a business from a user (from tip.json)