The dataset contains anonymous ratings(-10 to 10) provided by a total of 41,000 users. Train file contains 1.1 million ratings and in the test file the user needs to predict the ratings provided by the same set of users on a diffrent set of jokes. The complete text for all 139 jokes is also provided in a separate csv. Given the combination of user and joke, the task is to predict the rating given by that user to the joke in the test set

Public Private Split

Private split contains 60% of jokes from the test file rated by each user, while the public split contains the other 40% rated by the same user.

Evaluation Metric

The evaluation metric for this challenge is root mean squared error (RMSE)

Data

1. **Train link has ‘train.csv’ and ‘jokes.csv’**

**a. The train.csv file contains the user id, joke id and the rating**

|  |  |
| --- | --- |
| **Variable** | **Description** |
| id | Unique id |
| user\_id | ID for user |
| joke\_id | ID for joke |
| Rating | Rating given by the user to the joke |

**b. The jokes.csv contains the joke text**

|  |  |
| --- | --- |
| **Variable** | **Description** |
| joke\_id | ID for joke |
| joke\_text | Complete text for the joke |

**2. Test link has ‘test.csv’**

|  |  |
| --- | --- |
| **Variable** | **Description** |
| id | Unique id |
| user\_id | ID for user |
| joke\_id | ID for joke |