1. **Problem Definition**

In this project, We are given a dataset of records on drug review.In training dataset, it totally include 6999 samples. Validation dataset and testing dataset each include 1199 samples and 1799 samples.As shown in Table 1, the dataset provide patients’ rating from 1 to 5, higher rating indicate more positive opinion.Our work needs to find the relation between different attribute of the users’ comment and the rating.Then we need to use the training dataset to design and train the model, use validation dataset to evaluate our model and finally predict the rating of testing dataset.

|  |  |  |
| --- | --- | --- |
| Attribute | Type | Description |
| recordId | integer | record id of the review |
| drugName: | categorical | name of the drug |
| condition | categorical | condition to take the drug |
| reviewComment | text | review comment |
| date | date | review created date |
| usefulCount | integer | the number of users who find the review useful |
| sideEffects | categorical | level of side effects of the drug |
| rating | integer | 5 level patient rating |

Table 1. Dataset Attribute

1. **Data Analysis and Model Design**

First of all, I need to find out how different attribute affect the user rating.So I concat the training dataset and validation dataset to analysis the data.As it is shown in Figure 1, rating 5 is the most frequent rating that user gave, the following are rating 1 and rating 4. This indicate that if the drug satisfied users’ need, they prefer to give the rating five.Otherwise, they prefer to give the rating 1 rather than rating 2 or 3.

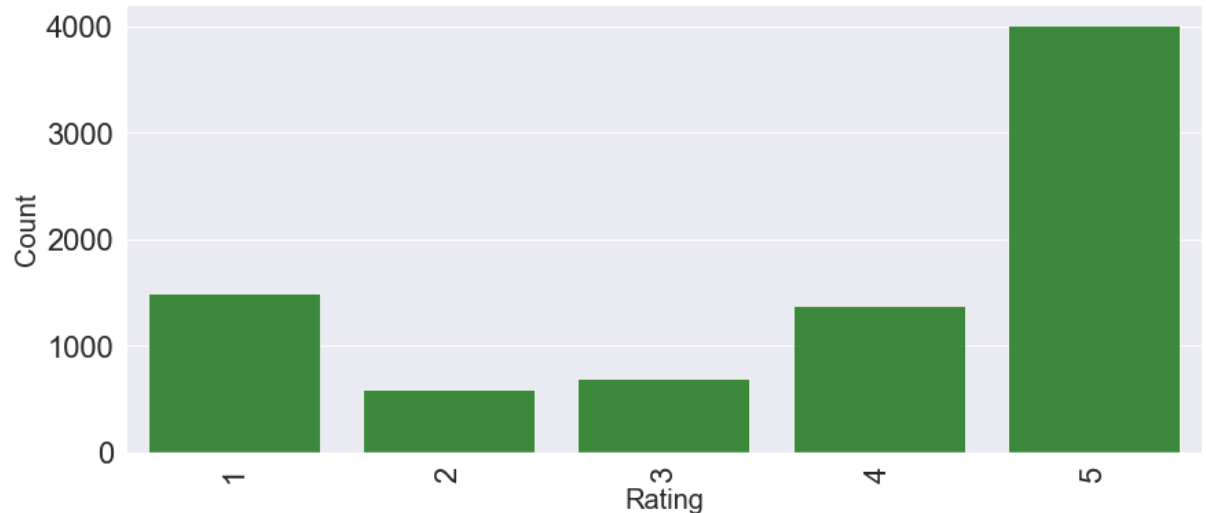


Figure 1. Distribution of numbers of different ratings.