CSE258 HW1

Yawen Zhao A53280596

1. The distribution of ratings in the dataset is as following:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Stars | 1 | 2 | 3 | 4 | 5 |
| Reviews | 4793 | 1569 | 3156 | 9859 | 129709 |

1. (only for CSE158)
2. The value of are

indicates the average value of ratings. indicates as if the review is verified, the ratings will increase by 5.04148265e-02. If the review is unverified, the ratings will not increase. indicates as if the review length is 1 character longer, the rating will decrease 1.24659895e-03. Conversely, if the review length is 1 character shorter, the rating will increase 1.24659895e-03

1. The value of are

indicates the average value of ratings. indicates as if the review is verified, the ratings will increase by . If the review is unverified, the ratings will not increase. Even though the coefficients represent the same feature, the model is being changed. As in question 3, the model of rating is represented not only by whether the rating is verified or not, but also represent by the length of the rating. Hence the model in question judge the rating in a different dimension and thus the coefficients are changes.

1. The model’s MSE on training data is 0.6557415620281802. MSE on testing data is 0.9714261885960409.
2. (only for CSE158)
3. The figure below shows the training and test error vary as a function of the training set size. The size of the training set makes a big difference of the testing performance. As the size of training set increases, the error rate of testing data increases. This is because

A close up of a map

Description automatically generated