5 RESULTS AND DISCUSSION

Due to the imbalanced rating distribution (e.g. more than 50% 5-score review) we have observed in section 2.2.1, the model would be inclined to focusing on 5-score review if we just use the top frequent words as features regardless of rating. Thus we choose to combine the top frequent words/phrases in 1-5 score review text to be our features. In section 3.4, we have compared different types of features involving review text, so we finally choose 1324 features (761 adjectives, 227 bigrams and 336 trigrams) plus a bias term to be used in our models. Then we use the models in section 4 to run test on test set.

5.1 Results of Linear/Ridge Regression

The results of linear/ridge regression models are shown as Table 5.1 below.

Table 5.1. The results of linear/ridge regression models

Model	Baseline	Linear Regression	Ridge Regression (λ=7)
MSE	1.7103	1.2185	1.2182

We can see from Table 5.1 that the result of using our text features outperforms that of baseline a lot. That is mainly due to that the length of review text is not as informative as the content in it as we have analyzed in section 2.2.2. And the ridge regression with regularizer λ=7 has almost the same performance with that of no regularizer.

5.2 Results of Logistic Regression

Since score has only the value of 1, 2, 3, 4 and 5, so we can try to treat this problem as a multi-class classification problem and solve it with classification method. So we try to use logistic regression to build our model, and the results are shown as Table 5.2.

Table 5.2. The results of logistic regression models

Model	Baseline	Logistic Regression (C=5.6)
MSE	1.7103	1.6984

The logistic regression model seems not to have a much better performance than the baseline. This is probably because we treat this problem as classification problem and that tends to increase the MSE.

5.3 Results of Gradient Boosting

In this section, we try to use gradient boosting regressor to build our models. Since gradient boosting is fairly robust to over-fitting so a large number of boosting stages will usually give the model better performance. Thus we choose the number of boosting stages to be 100. And we the maximum depth of the individual regression estimators is chosen to be 10. Other parameters are used by their default value. The results are shown as Table 5.3.

Table 5.3. The results of gradient boosting regressor models

Model	Baseline	Gradient Boosting
MSE	1.7103	1.1377

We can see that the MSE has exceeds that of the baseline by 33% and outperforms the linear regression model by 6.7%. Thus gradient boosting has a fairly good result.

5.4 Random Forest

Then we use random forest regressor to build our model and do test on test set. Similarly, we use the 1324 features chosen before, and run test with the number of trees in the forest (i.e. n_estimators) to be 100. The results are shown in Table 5.4 as below.

Table 5.4. The results of gradient boosting regressor models

Model	Baseline	Random Forest Regressor
MSE	1.7103	1.0585

The MSE of model built by random forest regressor is much lower than that of the model built by linear regression using the same features. This model also improves the baseline's MSE by around 38.1%, which is a fairly satisfying model.

5.5 Summary

In this part, we have tested different models on test data and also compared the results. Our random forest model has the best MSE performance, its performance exceeds that of the baseline essentially and also be greater than linear regression model by about 13% concerning MSE. This is mainly due to 2 reasons. Firstly, we introduce the numbers of top frequent adjectives as our features, and adjectives are naturally more informative since they are used to convey subjective evaluation in many cases. And also we introduce the bigram and trigram features to compensate the loss of information for splitting phrase into words. But other features are not as informative such as nouns, verbs and the length of reviews since they are less dependent on the rating of a review. Secondly, random forest has a good capability to resist overfitting due to its randomness and also has a good performance on imbalanced data due to its insensitivity to multicollinearity. This model has provided a method to do text mining and applies it on predicting people's rating.

For linear/ridge regression models, they have a relatively good performance although they are essentially simple models. This is mainly because they have used good features, but they cannot be outstanding due to the limit of the model's principle. The logistic regression's performance is very close to baseline because it's a classification method and we are using MSE to be the evaluation standard.

6 FUTURE WORK

Given more time, we can work on further improvement to our models in the following aspects:

- **Introducing dimension reduction before using features to train model**

The feature matrix may contain redundant information, so we can apply dimension reduction. Since the features are sparse, we can use SVD instead of PCA to do dimension reduction. This can eliminate some redundant information and reduce the computational cost.

- **Exploring more interesting features in the review text**

  There are still more interesting features in the review text need to be explored. For example, we can try adverbs, the number of all-uppercase words (which is a highly emotional feature) and so on.

- **Choose better parameters for our models**

  We can try parameters of our models more precisely. We can use better parameters such as the n_estimators in Random Forest model (which involves more time to compute) to get better performance.