Project Report
Review Rating Prediction Using Machine Learning

MT2023118 Janga Guru Pavani
MT2023179 Shreya Chavan

1 Introduction

1.1 Background

The project focuses on predicting review ratings based on various features such as user information, book details, review text, and additional metadata. The goal is to build a classification model that accurately estimates the rating given by users to a book.

1.2 Objective

Predict review ratings for books using machine learning techniques.Explore the impact of different features on the prediction accuracy.Provide insights into the factors influencing review ratings.

2 Data

The data-set contains the following columns:

- user_id: Id of the user.
- book_id: Id of the book.
- review_id: Id of the review.
- rating: Rating from 0 to 5.
- review_text: Review text provided by the user.
- date_added: Date the review was added.
- date_updated: Date the review was last updated.
- read_at: Date the user read the book.
- started_at: Date the user started reading the book.
- n_votes: Number of votes received on the review.
- n_comments: Number of comments on the review.

2.1 Train Data

Train data-set has all the above columns.It is used for training and validation of the model.Train data-set has 630000 data-points.The distribution of data according to ratings is show in Figure 1

2.2 Test Data

Test data-set has all the above columns except rating column.We have to predict ratings of the data-points using all the features of the Test data-set.Test data-set has 270000 data-points.

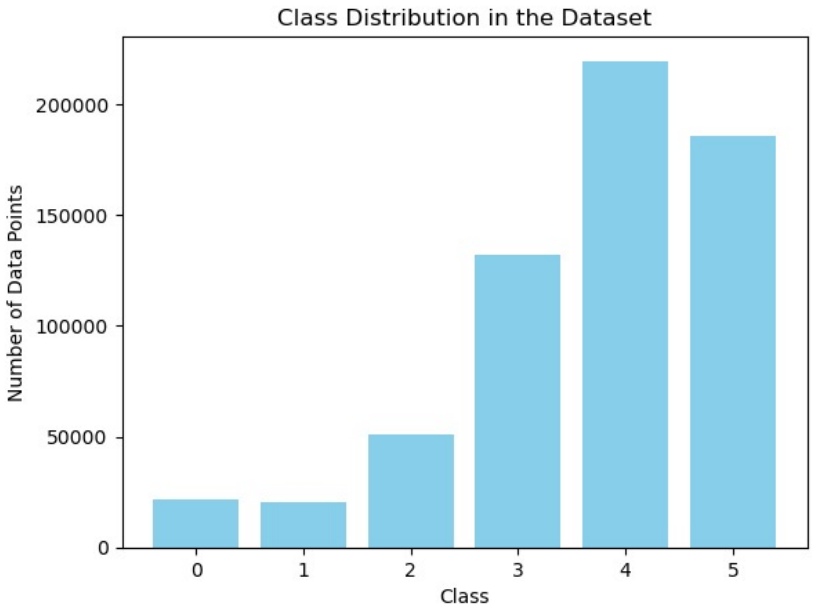

Figure 1: Class Distribution of Train Data.

3 Data Preprocessing

For data preprocessing step we filled null values with default values, used **tokenization**, **stop words removal** and **Lemmatization** techniques for text cleaning.We used Lemmatization instead of stemming as it does not give good meaning to the processed word but Lemmatization does.

After preprocessing we saved the data in respective csv files and used for modelling.

4 Modeling

4.1 Feature Engineering and Selection

We use TF_IDF vectorizer to transform the preprocessed review text to vectors.The other attributes are added to these vectors.These vectors are used for training the models and predicting the ratings.

4.2 Train-Validation Split

For training the model we are using the train data-set.We split the data into train and validation sets in 80:20 ratio.

4.3 Models

We used Logistic Regression and Naive Bayes Classifier for training with the entire data-set.We also tried with RandomForestClassifier and Support Vector Classifier(SVC) using partial data-set.

For training the model initially we used 'review_text' attribute and later we added other attributes which added additional information needed for modelling and validated the model using the validation data.

4.4 Metrics

The metrics used to find the accuracy of the model are **precision**,**recall** and **F1-score**.Precision is the number of true positives divided by the total number of predicted positives.It measures the accuracy of the positive predictions.Recall is the number of true positives divided by the total number of actual positives.It measures the ability of the model to capture all the positive instances.F1-score

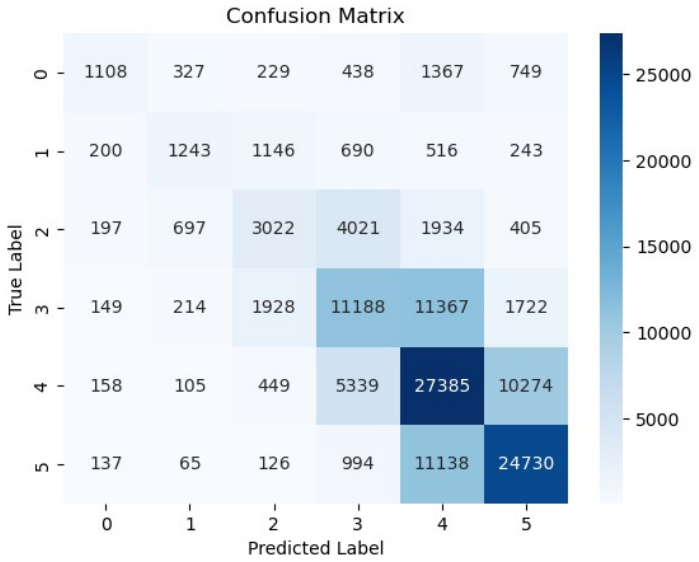

Figure 2: Confusion matrix of the predictions of Validation Data.

	precision	recall	f1-score	support
0	0.57	0.26	0.36	4218
1	0.47	0.31	0.37	4038
2	0.44	0.29	0.35	10276
3	0.49	0.42	0.45	26568
4	0.51	0.63	0.56	43710
5	0.65	0.66	0.66	37190
accuracy			0.55	126000
macro avg	0.52	0.43	0.46	126000
weighted avg	0.54	0.55	0.54	126000

Table 1: Metrics calculated for the validation data.

is the combination of precision and recall.It is the harmonic mean of precision and recall.It provides a balanced measure of precision and recall.

5 Results

In Logistic Regression using only 'review_text' attribute we got predictions with F1 score **0.53** for validation data.So we used other attributes like 'book_id','user_id','n_comments','n_votes' etc various combinations of the attributes in feature selection and got F1-scores in range of **0.52 to 0.54**.Using all the attributes except date related attributes the Model predicted ratings with F1-score **0.55**.The Table 1 shows the detailed accuracy for each class.Refer Figure 2 for the confusion matrix of the predictions of the validation data of the most accurate model.Using Naive Bayes Classifier the model gave F1-score of 0.43.For RandomForestClassifier and SVC we got around 0.35 to 0.4 F1-score for partial data.

6 Conclusion

Among all the models that we tried for predicting review ratings logistic Regression gave us the best results of F1 score **0.55**.