COMP30027 Assignment 2 Report

Aim

The aim of this project is to predict the ratings of books based on several features like titles, author, description of the book. There are three levels of ratings, 3, 4 and 5 for each book. The result may be interested by the bookshop or people who are interested buying new book. A rating of 3, 4 or 5 will be given when features of the book is provided, this may help the bookshop to decide whether to purchase the book or not.

Datasets

The dataset is collected from Goodreads, this platform is used to rate books and write reviews for books. There are 23063 pieces of data with 10 columns of attributes.

Each piece of data describes the basic information about the book, and it introduces:

- Name: name of the book
- Author: author of the book
- PublishYear: publish year of the book.
- PublishMonth: publish year of the book.
- PublishDay: publish day of the book.
- Publisher: publisher of the book
- Language: language the book is written in
- PagesNumber: the number of book pages
- Description: brief knowledge of the book
- rating_label: rating of the book

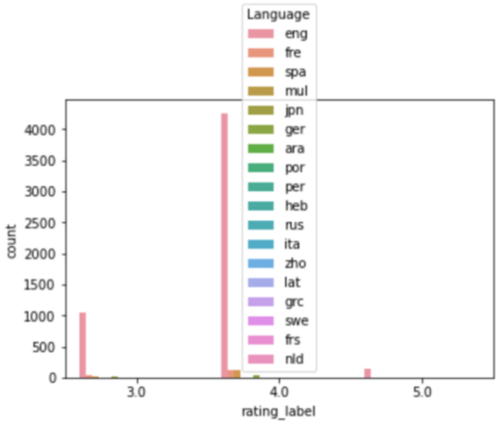
The rating_label of the book is the response variable for our investigation.

Data wrangling

The data have firstly been preprocessed, firstly convert all columns that contains String attributes to lower case, this decreases the vocabulary list by a large factor.

The data quality is then measured. There are 5 duplicates data pieces that are deleted and there are a few missing values in publisher column and a lot of missing values in language column. The missing values in publisher column is simply ignored, when trying to add the language of the book to the data set, natural language processing step can be used to predict the type of language from description and book name. After further infestation about the dataset, 93.0% of the books are written in English, we were trying to replace all missing language books to English book. However, after doing that, 98.2% of the books are written in English which will cause distortion of information. Instead, we just ignore the fact that there are missing values in “Language” column. Figure 1 is shown below.

Figure 1 ratings for books with different languages



Analysis Methods

Methods

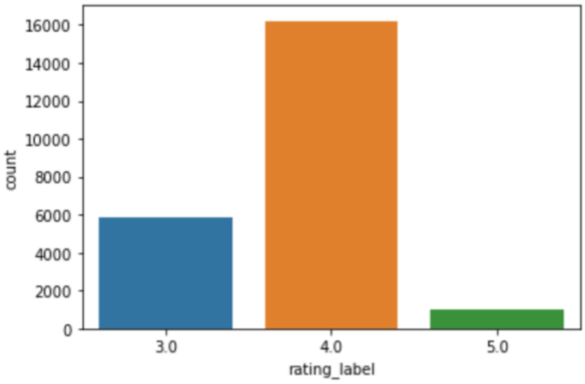
We deliberately chose to use Chi-square test feature selection techniques to get which factors are important determinants of rating label, “Name” and “Description” is rejected in Chi-squared test. The chi-squared test is a useful tool for feature selection in categorical data, but it may not be appropriate for evaluating the importance of “Name” and "Description” features in this project. since this investigation deals with ordinal data (rating label), we use multinomial Naïve Bayes, logistic regression, Linear SVM and KNN.

Train-test split

Before further analysis, the data is split into training and development sets (0.9/ 0.1 split). he training set is used as a technique for evaluating the performance of a machine learning algorithm, whereas the development set is used to assess the fit and evaluate the models.

Preliminary Analysis
Firstly, we observed that around 70% of the rating label is 4 showing in figure 2 below, this highly unbalanced label, this will cause bias towards majority class, poor generalization to minority class, class weighting or resampling technique can be used to reduce the effect.

Figure 2 Number of books for different ratings



To gain a preliminary understanding of which book-related factors have the greatest impact on the rating of each book. Heat map between each variable with Pearson correlation and selected significant factors through the Chi-square Feature Selection. We observed that related to the Pages Numbers and Publisher showed deeper colors and Publish Year showed lighter colors. On the heat map, the larger values contained in a data matrix are represented as deeper colors, so these three features seemed to hold correlations with the rating label. Relationship between page numbers and rating labels is furthermore investigated. We found that the mean page numbers is increasing with respect to the rating labels. However, there are several extremely larger page numbers in the data set therefore top 1% outliers are removed and boxplot is drawn. The mean page numbers for different rating labels are about the same, however, the 75% quantile is increasing with rating labels which is a sign saying high rating books tends to have more page numbers. This is due to the fact that the label is highly unbalanced. Pearson correlation of 0.079 is relatively weak between page number which implies there is not a strong positive linear relationship between page numbers and rating labels.

Figure 3. Heat map with Pearson correlation for the dataset

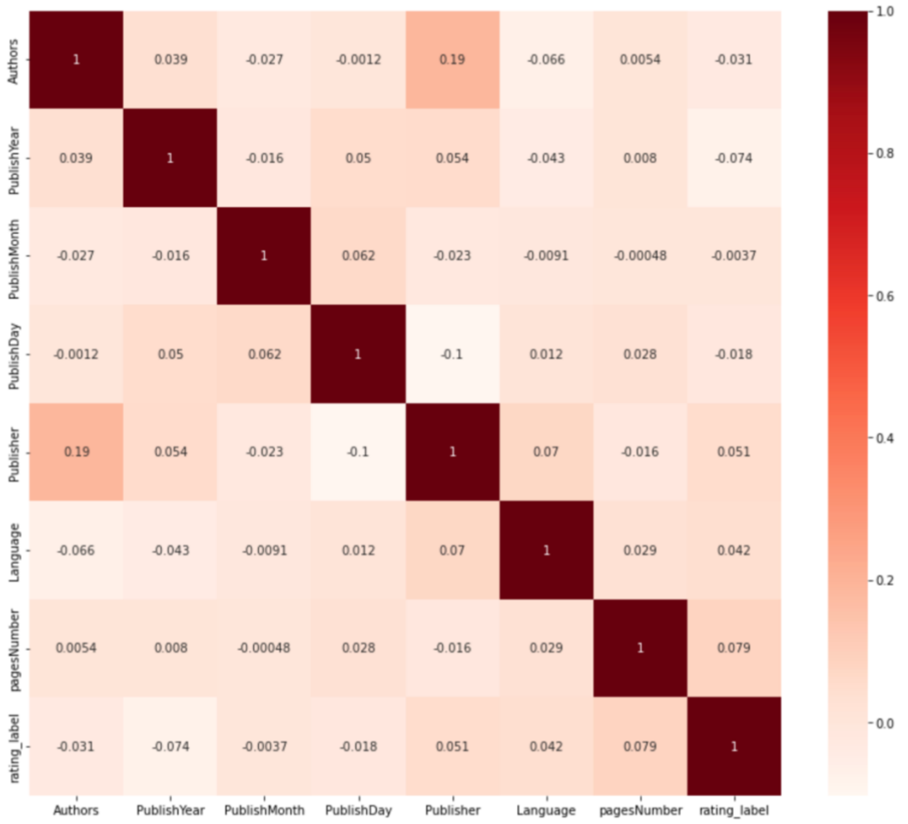
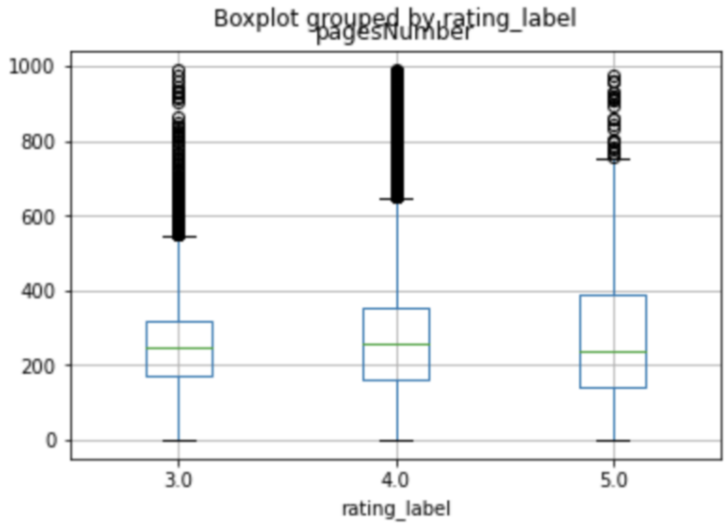


Figure 4 Boxplot of page numbers for different rating label books



To further select the significant factors, Chi-square test Feature Selection is used. We were able to observe the p-value with the Chi-square test to determine the relationship between the testing feature and the response variable – rating_label. 7 features are selected and they are “Author”, “PublisYear”, “PublishMonth”, “PublishDay”, “Publisher”, “Language”, “PagesNumber”. Name and Description is rejected as this feature selection is unable to recognize the importance of texts.

Modelling

As the rating_label was the response variable in our investigation, it held a ternary classification problem. Four models are fitted. Overall, four models produce similar performance around 70 percent accuracy.

Table 1. Parameters for four different models

Models	Accuracy (%)
Multinomial Naïve Bayes	70.86
Logistic Regression	70.16
Linear SVM	70.26
KNN	68.34

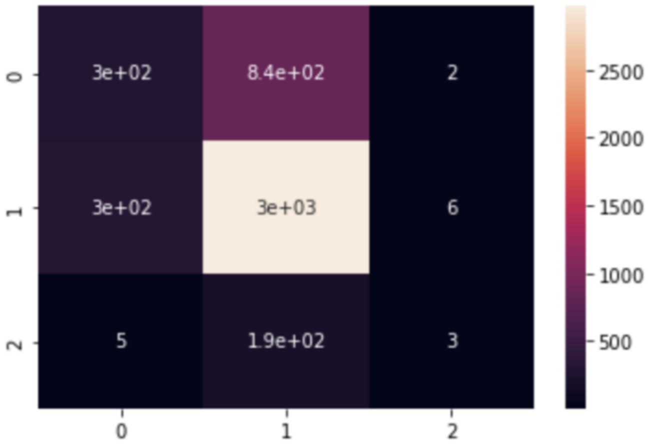
Discussion

From the previous analysis, Multinomial Naïve Bayes is the best model for the dataset, the accuracy is 70.86% whereas KNN has the lowest accuracy. We furthermore discover the performance of multinomial Naïve Bayes and confusion matrix is drawn. The classification report is shown in table 2 and heatmap is shown in figure 5 below.

Table 2 classification report of multinomial Naïve Bayes

Rating_label	Precision	Recall	F1-score
3	0.49	0.26	0.34
4	0.74	0.91	0.82
5	0.27	0.2	0.3

Figure 5 Heatmap of confusion matrix of MNB



Compared to the performance for rating 3 and 5, rating 4 has relatively high precision, recall and f1-score. This is again since the label is highly unbalanced, compare the recall and precision for rating 4, it has relatively high recall than precision. This implies that the model does not require much evidence to say the book will have rating of 4. Rating 3 and

5 books have higher precision than recall implies that the model needs lots of evidence to say “positive”. In the confusion matrix, the leading entry is supposed to have light colors representing label is correctly predicted, however, only when the actual type is rating 4 and the predicted type is rating 4 has light colors, this implies that most books with true rating label of 4 will be correctly predicted as rating 4. Strange thing happened here, a lot of books that have rating of 3 are mistakenly predicted as they are rating 4. This is because 70 percent of the book have rating of 4 and only 25 percent of book have rating of 3. There is a poor generalization to the minority class.

Then we tried to identify the words that may lead to a higher rating in “Description”. We then selected the most important 20 words and result is:

['br', 'bible', 'god', 'volume', 'novel', 'photographs', 'spiritual', 'translation', 'internet', 'poems', 'work', 'web', 'reference', 'christ', 'poetry', 'divine', 'comprehensive', 'online', 'published', 'study']

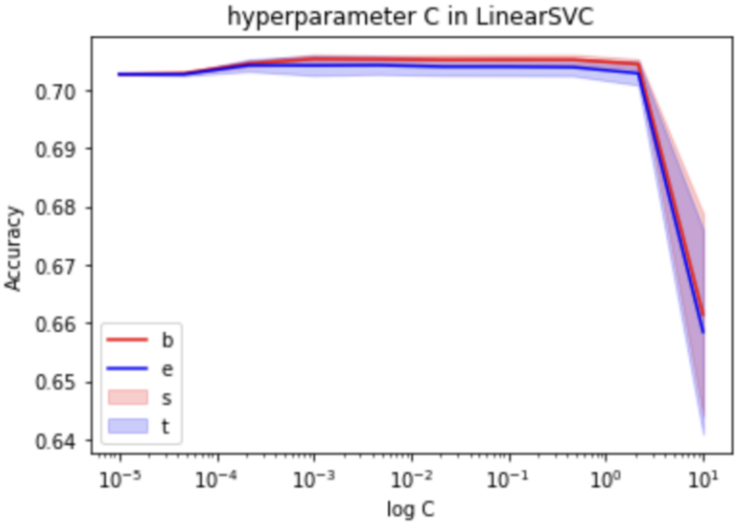
Books with Description key words like “bible”, “spiritual”, “poems”, “christ”, “poetry”, “divine” are more likely to have a higher rating. This may imply that many of the readers are religious and religious readers are more likely to rate higher.

We then did a Logistic Regression, In logistic regression, the dependent variable is binary, taking values of 0 or 1. The goal is to estimate the probability of the positive class (Y) given a set of independent features. The output of logistic regression is a probability score between 0 and 1. We take one class as pivot and build a regression model for every other class cj. We treat class cj as class Y and the pivot class as class N. During the training process, logistic regression learns the optimal coefficients for the independent variables that maximize the likelihood of observing the given data. These coefficients represent the importance or contribution of each feature in determining the class probability. Finally, we average the cross validation score and get an accuracy of 70.16% which is a little worse than Multinomial Naïve Bayes, this is due to the unbalance of label values.

Linear Support Vector Machine (SVM) is used, SVM tries to find a linear decision boundary that maximally separates the classes in the feature space (Description in this case). The decision boundary is defined by a hyperplane, which is a flat subspace in the feature space. The SVM algorithm aims to find the hyperplane that achieves the largest margin, which is the maximum distance between the hyperplane and the closest data points from each class.
We did parameter Tuning to find the best hyperparameter C in linear SVC at around 10e-3, this value gives the highest accuracy for linear Support Vector Machine. Hyperparamter C is shown in figure 6 below.
After fitting SVM model to the dataset, we average the cross-validation training accuracy score and get a final result of 70.26 which is better than Logistic regression. It optimizes a cost function that balances maximizing the margin and minimizing the classification errors by solving a convex optimization problem to find the optimal hyperplane that achieves the

best separation between the classes. However, as the label is unbalanced, the performance of SVM is about the same with Logistic regression and Multinomial Naïve Bayes.

Figure 6 Hyperparamter C in linearSVC



Finally, KNN is used to predict the label and it achieves an accuracy of around 68.34 The Author column is very sparse, 23063 pieces data with over 16000 authors, which makes the K nearest neighbor algorithm harder to achieve high accuracy. We firstly tried k = 3 and find accuracy is only around 63% which is lower than expectation. After adding k to around 100, the accuracy stays at around 68 percent. This is because the data is sparse in the space, there are over 4000 different Publishers and over 16000 authors, and the data is unbalanced for column Language, over 93 percent of the language is written in English. This reduces the accuracy of KNN model.

Conclusion

Cross validation is used to evaluate the performance of the machine learning algorithm which is more accurate than only using single train-test split. It reduces the randomness for extreme cases.

As our aim is to predict the rating of the book by firstly finding the related factors which have a relationship with “rating_label”. Based on the general knowledge of the review of books, readers have the habit to rate 4 when they think the book is interesting but not perfect. That might be the reason why there are so many rating label of 4 than rating label of 3 and 5. The nature of having too many rating label of 4 causes the unbalance of data, this makes prediction harder. As the percentage of rating label of 4 is around 70%, even 0R can get an accuracy of 70% by predicting everything as rating label of 4. class weighting, resampling technique can increase the performance of models. Decrease the number of rating label of 4 data pieces may also increase the performance of models as models has bias for majority classes.

The Language column is also highly unbalanced, over 93% of the book is written in English, this causes this “Language” feature “Loss of important information, models focuses too much on languages like English or French, too few books are written in other languages, thus models will not learn enough about other languages.

Moreover, there are too many missing values in column Language, we lose too many data pieces if we remove all the missing data, in this project, we simply just ignore the missing values. In the future, better method like predicting missing “Language” by using natural language processing can be used.

Reference:

COMP20008 element of data processing, project 2
https://www.machinelearningplus.com/pandas/pandas-dropna-how-to-drop-missing-values/
https://stackoverflow.com/questions/8420143/valueerror-could-not-convert-string-to-float-id
https://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html
https://pandas.pydata.org/docs/reference/api/pandas.Series.str.lower.htm
https://www.ibm.com/topics/logistic-regression#:~:text=Resources-,What%20is%20logistic%20regression%3F,given%20dataset%20of%20independent%20variables.