Machine Learning Assignment 2 Reflection

Ngoc Duy Tran – 1156724

1. Process

The whole process of producing this report can be followed by the Jupyter notebooks zipped in the submission part.

Firstly, data visualization gives a better understanding about the dataset, data distribution. After having a better insight into dataset, data wrangling methods such as scaling, imputation are applied for better data quality results. Regarding text analysis, through some experiments with baselines, this research highlights that TF-IDF Vectorizer shows a better performance than Count-Vectorizer, which is consistent with previous findings.

During the data pre-processing part, *book descriptions* is reported to contain valuable information for predicting the response variable – book rating, which signals a highly predictive feature. This, combined with certain disadvantages of traditional count-based text vectorization, , might imply pre-trained deep learning models on huge dataset for NLP tasks utilizing word embeddings, might show a relatively good performance on this book rating dataset.

2 machine learning models (LGBMClassifier + Pre-trained Small-BERT) are evaluated on the validation and test sets. By analyzing errors from both models, a common difficulty from both models is too many predictions are targeted at the major class, while the recall for the other 2 minor classes is relatively low. Based on this difficulty, the project proposes several methods such as adjusting class weights, utilizing SMOTE, Focal Loss for Pre-trained BERT.

2. Satisfaction and further improvements

Overall, I feel relatively satisfied with the report because structurally, it sticks closely to a standard conference paper. Although there are only 2 models in the modelling part, I believe that I have managed to develop critical thinking and careful elaboration for the discussion and error analysis part. I have carefully analyzed the error, and also devised certain strategies to deal with errors while trying to maintain the overall accuracy (SMOTE, Focal Loss). Model evaluation on validation set is also rigorous as it covers both summarative figures (Table) and detailed analysis (Confusion Matrix Visualization) for deeper understanding.

However, one thing that needs to improve is on modelling because the current rank of the models is not as high as expected. If I had had more access to computational resources – GPU, I could have experimented more models: Multilingual-BERT, RoberTa, which potentially yields higher scores. Further evaluation metrics can also be applied such as ROC-AUC for further analysis.