COMP300027 Machine Learning
Assignment 2: Reflection

Sammi Li 1271851

Word Count: 394

My first task was to determine how to pre-process the textual and categorical data. I used the Doc2vec vectors given over tokenization and bag of words, as it can capture the semantic meaning of words. I then changed the features, “Publisher” and “Language” into numerical representations of their categories. Then, I chose to use the Pointwise Mutual Information (PMI) score between features and rating label to select the top 20 features to funnel into the classification models. This was done to accelerate the model building time and improve performance as redundant and irrelevant features are not considered. The next step I took was to create simple models of different regression and classification models to test the baseline performance of all models and choose 3 models to focus on. Initially, I chose linear regression, random forest and support vector machines (SVM). However, I found that logistic regression would be a more suitable “linear” model for book rating classification than linear regression. I also found that SVM is a very time-inefficient model due to the large datasets and extensive combinations of hyperparameters, taking over 4 hours for the model to learn and fit into the dataset. Therefore, I chose to remove testing of the SVM model due to time constraints and productivity of the project. Thus, I concluded that a Naive Bayes model was to be constructed as the baseline of comparison as it is both simple and commonly used. I then conducted hyperparameter tuning on the logistic regression and random forest classification models through grid search and trial and error of parameters given to the grid search function. This section required plentiful trial and error where I found it difficult to determine which parameters to change and which had the most impact towards the performance of the model. Due to the extensive execution times of these models, it was difficult to test different parameters quickly, so I had to choose widely to have more productive use of time. In my stage 1 deliverables, I am satisfied with the choice of diagrams used, and how I approached the text pre-processing and feature selection. I do wish that I had considered more complex models like SVM and neural networks. I also think that my solutions towards the error analysis section could be implemented in a more in-depth manner, which may lead to better performance and model tuning.