Feedback 1

1. Report Summary

The main content of report is split into 3 parts: Data Pre-processing, Model Training Results and Discussion. In **pre-processing**, the author develops 2 schemes for categorical feature selection (intuition and chi-squared based). For text features, the author incorporates BoW for text feature transformation and argues that semantics are not too important for this task. Feature selection implementing Chi-square test is then applied to reduce dimensionality for better generalization. The author also tunes the number of features for 4 machine learning models based on different strategies (hold-out, cross-validation) for generalization.

In **Model Construction**, 4 machine learning algorithms are experimented, including Logistic Regression, K-NN, Multinomial Naïve Bayes, and Stacking. Results are delivered in table/figures.

In **Discussion** part, the author critically analyses the learning curve, and provides crucial reasons behind, and conclude that chi-squared based features work better.

2. Effective parts

Overall, the author's data pipeline is very well-prepared, in which it covers essential steps for ML models such as feature selection. The error analysis based on learning curves, to me, is very detailed and rigorous as it covers possible reasons and makes assumptions to understand the declining performance when the number of features increases. Visualizations are easy to interpret, which supports the arguments presented. Although machine learning models are simple, the work behind is well-prepared.

3. Improvement

Despite a high-quality report technically, further improvements can be applied regarding presentation. Minor issues with abbreviation (wouldn't) should be avoided, and visualizations are a bit overwhelming (34 figures), and far from **Discussion**, which might cause certain difficulties for readers.

Regarding technical issues, I think the data-preprocessing step can be improved by introducing a HTML tag removal functionality, which might improve the overall performance.

Despite the above minor issues, this report is of high-quality.

Feedback 2

1. Report Summary

The report introduces the background of the task – book rating prediction. The main content of report can be divided into 3 parts: Data Pre-processing, Model Training and Evaluation. In the **pre-processing** part, the author incorporates the use of basic text-preprocessing steps such as lowercasing, tokenization; along with TF-IDF Vectorizer for feature transformation for machine learning algorithms input. After this step, feature selection implementing Chi-square test is also applied to reduce dimensionality for better generalization.

In **Model Construction** part, 3 distinct machine learning algorithms are experimented, including DecisionTree, KNN and MLP (Multi-Layer Perceptron) for analytical purposes. 3 models' implementation is based on sklearn module.

In **Evaluation** part, the author reports all 3 machine learning models achieve relatively competitive results, above 60%, and conclude that neural-based networks achieve the highest performance regarding overall accuracy on test-set

2. Effective parts

Overall, the author's data pipeline is relatively well-prepared, in which it covers data pre-processing steps such as text-preprocessing and feature selection. Moreover, the author also highlights the advantages of TF-IDF Vectorizer instead of Count-Vectorizer, which I assume a reasonable step. Moreover, the author also made a comparison between 3 models and also draw a conclusion from this experiment.

3. Improvement

In addition to part 2, I think this project lacks the reproducibility elements, in which model configurations for the 3 models are not specified. The author's model could have been more detailed and rigorous by partitioning the training set into 2 subsets: training and validation, from which error analysis is conducted. Therefore, the current evaluation part is relatively restricted.

Structurally, I think the report does not follow closely to the template or any conference templates.

Overall, the report covers all the basic steps of a machine learning pipeline, yet further improvements can be applied for better quality.