Comparing Scalable Multiclass Classification Models to Rate Amazon Books

Hsiu-Ping Lin, *Min-goo Kang, Jongwook Woo
Department of Information Systems, California State University, Los Angeles
*Department of Computer Information Systems, Hanshin University, Korea
{hlin54, jwoo5}@calstatela.edu, kangmg@hs.ac.kr

Abstract

This paper aims to predict the book's ratings using the large-scale dataset of Amazon books. The dataset is too large to handle with the traditional approach, so we adopt the Big Data platform, PySpark, in the Amazon EMR cluster that processes large-scale data. We build prediction models as multiclass classifiers using various PySpark algorithms – Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), and Multi-Layer Perceptron Neural Network (MLP-NN). So, we can predict book ratings in five and binary classes. We compared Multiclass Classifiers' accuracies, especially Precision, which has the smallest False Positive. As a result, we observed that the LR model exhibits the highest precision in binary and multiclass classification models and boasts the fastest computing time. Besides, MLP-NN demonstrates good performance, although it requires more time for model construction than LR. RF and DT show lower performance and demand more computing time than LR and MLP-NN. Furthermore, binary classification models have better performance in precision than multiclass ones.

Keywords: Big Data, Spark, Scalable Computing, Neural Network, Multiclass Classification, Rating

1. Introduction

Amazon has emerged as one of the world's most successful e-commerce enterprises. Its book-selling business has been producing data related to curated lists of bestsellers, popular books, and a recommendation system. Reviews text of the data are generated for the respective products when customers provide feedback and ratings on the website. It would be helpful to have a prediction model to recommend the right books to customers. The challenge for this predictive analysis is that first, the data size is too big to use the traditional computing systems, and second, review data requires text processing and vectorization, which requires more computing time. So, this paper employs Big Data platform to build PySpark Machine Learning Models to predict book ratings efficiently using the Amazon book review dataset from Kaggle. The PySpark Big Data cluster comprises multiple nodes for distributed parallel computing, which also attain scalability.

2. Related Work

Chen et al. explore machine learning models to predict Amazon product sentiment analysis. It aims to create an accurate model that can handle diverse sentiments expressed in reviews. The study evaluates NLP models (Light GBM, CatBoost, deep learning). The article discusses potential applications for sentiment analysis, concluding with results, key findings, and recommendations for future research [1]. Norinder et al. combine deep learning and conformal prediction for accurate sentiment analysis of Amazon product reviews across 12 categories. The paper emphasizes the approach's generalizability and its ability to handle imbalanced sentiment classes in Amazon reviews [2]. Our paper built LR, DT, RF, and MLP-NN models to predict ratings of Amazon Books in Big Data platform. We compare the computing time and accuracy of the models.

3. Data and Architecture

The dataset, obtained from Kaggle [3], comprises two files: *Books_rating.csv* and *Books_data.csv,* with a size of 2.9 GB. *Books_rating.csv* is a substantial file, containing information on 3 million book reviews, including user details for each review, along with columns: [ID, Title, Price, Profile name, Review Summary, Review Text, Review Helpfulness, Review Class]. *Books_data.csv* provides additional information, featuring details on genres, authors, cover designs, and descriptions, with columns: [Title, Description, Authors, Publisher, Categories].

It is not easy to handle data of 2.9 GB using traditional systems. So, we adopt the Big Data platform within the Amazon EMR cluster to process large-scale datasets. Table 1 shows the specifications of the cluster with Hadoop and Spark services.

Table 1. EMR cluster Specification

Number of nodes	3
Hadoop Cluster Version	3.2.1
Spark	3.1.1
CPU speed	1.995 GHz, 8 cores
Memory	536.4 G

4. Machine Learning Models

It is vital to recommend customers with higher-rated books. Our objective is to predict the rating classes of Amazon books using various data features.

We constructed prediction models as binary classifiers employing various PySpark algorithms, including Logistic Regression (LR), Random Forest (RF), Decision Tree (DT), and Multi-Layer Perceptron Neural Network (MLP-NN). Achieving the accuracy of each model, the smallest number of False Positives is more important than minimizing False Negatives. In other words, *Precision* takes precedence over *Recall*. Amazon book ratings range from 1 to 5, with 5 being the highest score. In this structure, binary classification is also possible. In binary classification, we have assigned Classes 1, 2, and 3 to the label 0, and Classes 4 and 5 to the label 1.

To address missing values in the numeric features, we replaced them with their respective average values. For the text feature, we removed the corresponding rows. To handle the various data types of features, we processed the numeric features using MinMaxScaler, while the text feature underwent several NLP preprocessing steps, including *Tokenizer*, *StopWordsRemover*, *CountVectorizer*, and *IDF*. Additionally, we utilized *CrossValidation* techniques to generalize the models with hyperparameter tuning.

Table 2 represents each model's computation time (CT), Area Under Curve (AUC), Precision, and Recall.

Table 2. Performance of Binary Classification

Algorithm	CT (min)	AUC	Precision	Recall
LR	16	.87	.86	.86
RF	486	.56	.84	.80
DT	756	.64	.83	.82
MLP-NN	222	.86	.85	.86

The study aims to identify a model that can recommend higher-rated books to customers, focusing on achieving highest *Precision* with the shortest computing time. **Fig. 1** represents the confusion matrix of LR. **Table 2** reveals that LR performed the highest *Precision*.

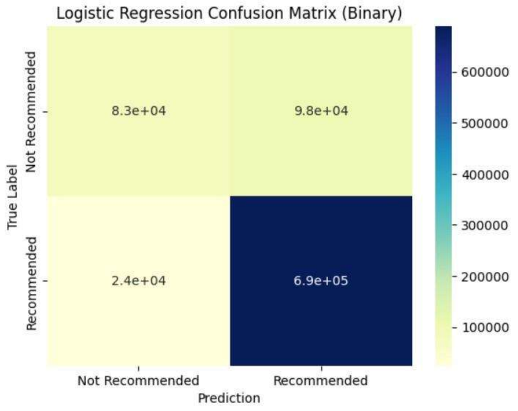

Fig 1. Confusion Matrix of Binary LR

LR model demonstrates the highest *AUC* (87%) and Precision (86%). It also requires a significantly shorter time (16 min) than other models. Beyond traditional machine learning models, the MLP-NN shows a high *AUC* (86%) and Precision (85%); however, it requires 14 times much longer running time than LR. As RF and DT exhibit relatively acceptable *Precision* and *Recall*, their *AUC* scores are 56% and 64%, respectively, which is notably unreliable compared to LR and MLP-NN.

Table 3. Performance of Multiclass Classification

Algorithm	LR	RF	DT	MLP-NN
Computing Time (min)	19	492	1458	204
Precision				
Class 1	.59	.93	.65	.55
Class 2	.42	-	.46	.38
Class 3	.46	-	.62	.44
Class 4	.47	.62	.44	.45
Class 5	.71	.60	.63	.68
Recall				
Class 1	.43	.005	.12	.39
Class 2	.18	0	.05	.11
Class 3	.23	0	.05	.18
Class 4	.21	.003	.11	.14
Class 5	.94	.999	.97	.95

By adhering to the labels in the initial dataset, which range from score 1 to score 5, we have also presented the performance of multi-classification for each algorithm in Table 3. We observe that multi-classification performs less effectively than binary classification due to five classes, as expected.

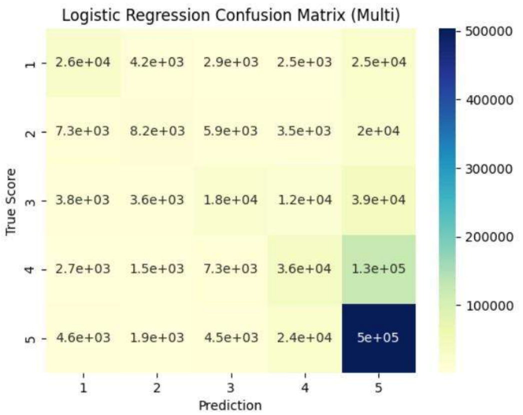

Fig 2. Confusion Matrix of Multiclass LR

LR outperforms others in *Precision* for Class 5 (71%), and has the shortest running time (19 minutes), as shown in **Table 3**. Although the MLP-NN demonstrates close *Precision* (68%), it requires 11 times longer running time (204 minutes) than LR. Additionally, RF predicts Class 1 with the highest *Precision* (93%), which is highly unfavorable to recommend a book.

5. Conclusion

The paper is to build models that can accurately predict Amazon book ratings and recommend a book to customers. The Amazon book data set is large-scale data, 2.9 GB, which does not allow the traditional systems to process complicated machine learning computation, so we adopt a Big Data platform, Spark cluster, using Amazon EMR service.

We developed binary and multiclass classification models - LR, RF, DT, and MLP-NN - for predicting Amazon book ratings. We compared the models with the accuracy and computing time as they process a large-scale data set. LR demonstrated the highest *Precision* and the shortest computing time. MLP-NN model also performed close Precision to LR, but necessitated 10 – 14 times longer computation time than LR.

References

[1] A. Chen, J. Walsh, M. MacDonald, N. Chu, R. Ahmed, and S. Rao, “Amazon Review Rating Prediction with NLP”, 2021. Retrieved from https://medium.com/data-science-lab-spring-2021/amazon-review-rating-prediction-with-nlp-28a4acdd4352

[2] U. Norinder, & P. Norinder, “Predicting Amazon customer reviews with deep confidence using deep learning and conformal prediction”, Journal of Management Analytics, Volume 9 (1), 2022.

[3] Mohamed, B. Amazon Books Reviews, 2023. Retrieved from https://www.kaggle.com/datasets/mohamedbakhet/amazon-books-reviews