Journal of Intelligent System and Applied Data Science (JISADS)

Journal homepage : https://www.jisads.com

ISSN (2974-9840) Online

ENHANCING UNIVERSITY ELECTRONIC BOOK ACQUISITION STRATEGY USING A DEEP FOREST FUSION APPROACH

Sanjai Kumar. T[1] *Mohammad Sameer Aloun*[2*]

[1]*Periyar Maniammai Institute of Science and Technology, Thanjavur, Tamil Nadu, India*

[2]*Faculty of Science and Information Technology Irbid National University*

Abstract

This study presents a new approach, LHGCAT-XDF, to improve the efficiency of electronic book procurement in university settings by combining the strengths of the LightGBM and CatBoost algorithms. This innovative model benefits from the LightGBM's minimal memory usage and CatBoost's reduced time complexity. Through testing, it's shown that LHGCAT-XDF surpasses standard machine learning models in overall effectiveness, successfully addressing the shortcomings of conventional procurement strategies in terms of accuracy and efficiency. Thus, it offers dependable guidance for the selection of electronic books in university libraries.

Keywords : Machine learning, LightGBM, CRTF, CatBoost

1. INTRUDUCTION

In the rapidly evolving information society and with the widespread adoption of mobile devices, the borrowing volume of physical books has been on a consistent decline. Concurrently, the demand for electronic books among readers has been increasing. This trend not only fosters a growing demand for a variety of electronic books but also sets higher standards for their quality and the services provided.

The primary goal of constructing university libraries is to meet readers' demands for more accessible and diverse academic resources, thereby enhancing service quality to align with the trends of the information age. This entails not just updates to physical infrastructure and spatial optimization but also improvements in the volume and quality of literature resources and an overall enhancement of library services [1]. In this transformation, libraries must evolve from traditional service models to smarter, more reader-centric approaches. However, most domestic university libraries still adhere to conventional book procurement strategies, relying on annual budgets, the experience of purchasers, recommendations from faculty and students, and suggestions from vendors to compile their procurement lists [2].

While some university libraries have begun to employ information technology to develop decision-support systems for book procurement, most systems still base their purchasing decisions on existing collection catalogs, borrowing data, and reader information, using statistical analysis [3]. Given the dynamic and uncertain nature of this data, understanding readers' reading needs and purchasing books that best meet these needs within a limited budget remains a key challenge in the book procurement process.

To enhance the efficiency and quality of electronic book procurement in university libraries and to explore the complex and variable relationship between book attributes and reader demands, this article adopts a hybrid deep forest model for predicting electronic book procurement in university libraries. This model not only significantly improves prediction accuracy compared to traditional machine learning models but also reduces the time complexity of model predictions and the difficulty of tuning hyperparameters, making it a more accurate and efficient algorithm for the field of book procurement prediction.

2. DEEP FOREST MODEL

The Deep Forest model, introduced by Zhou Zhi-Hua and Feng Jie in 2019, represents an ensemble method based on decision trees, falling under the umbrella of decision tree ensemble techniques [4]. Unlike Deep Neural Networks (DNNs), Deep Forest showcases superior competitiveness by requiring fewer hyperparameter adjustments, thereby reducing the time cost associated with hyperparameter tuning. It adapts well to datasets of various sizes and exhibits excellent generalization capabilities. These advantages have led to its wide application across different fields, affirming its robustness in classification and prediction tasks.

The Deep Forest model consists of two main components: Multi-Grained Scanning and Cascade Forest. Multi-Grained Scanning aims to analyze input features to unearth the sequential relationships between them. This process involves scanning the input feature vector with sliding windows of various lengths, generating multiple k-dimensional feature fragments [5]. These fragments are then fed into Random Forest (RF) and Completely Random Tree Forest (CRTF) models, with their class probability vectors, concatenated to form a transformed feature vector for the Cascade Forest input.

Cascade Forest, structured in multiple levels, each contains several ensemble learning classifiers, such as decision tree forests, XGBoost, LightGBM, or CatBoost. This hierarchical organization aims to build a stronger ensemble with better generalization performance. The model's design allows flexible feature learning and combination, with k-fold cross-validation employed in training each forest to prevent overfitting. The cascade structure dynamically adjusts the number of levels based on the training process, enhancing model complexity adaptability and training loss control.

To further enhance the performance of Deep Forest on smaller datasets, this paper introduces optimizations through LightGBM and CatBoost algorithms, simplifying the Multi-Grained Scanning structure and optimizing the number of random forests [6],. LightGBM reduces the number of data instances with minor gradients by utilizing one-sided gradient sampling, thus saving time and space [7]. CatBoost, through an optimized gradient boosting method and combining symmetric tree models with feature quantile metrics, simplifies model training and reduces data preprocessing complexity, offering an efficient and precise solution for electronic book procurement prediction, particularly with text-type electronic book attributes [8].

3. DEVELOP AN E-BOOK FORECASTING MODEL USING AN OPTIMIZED DEEP FOREST ALGORITHM.

Utilizing the past five years of access records from S Academy Library as a foundation, this study constructs a precise model for predicting e-book procurement. Initial steps involve preprocessing the gathered data and employing value indicators to sift through e-book interview decision influencers. The BM25 algorithm [9] facilitates the engineering of text features, with the refined samples then applied to develop the prediction model using the LHGCAT-XDF approach, as illustrated in Figure 1.

Feature Analysis

Diverse decision-making elements influence e-book interviewers across university libraries, with varying degrees of importance attached to each. An exhaustive analysis, incorporating practical insights from the e-book acquisition processes in numerous university libraries and academic perspectives, leads to a detailed summation of the current influential factors in library acquisition decisions. Following this, characteristic variables are identified and quantified via information gain to aid in constructing the subsequent e-book prediction model [10].

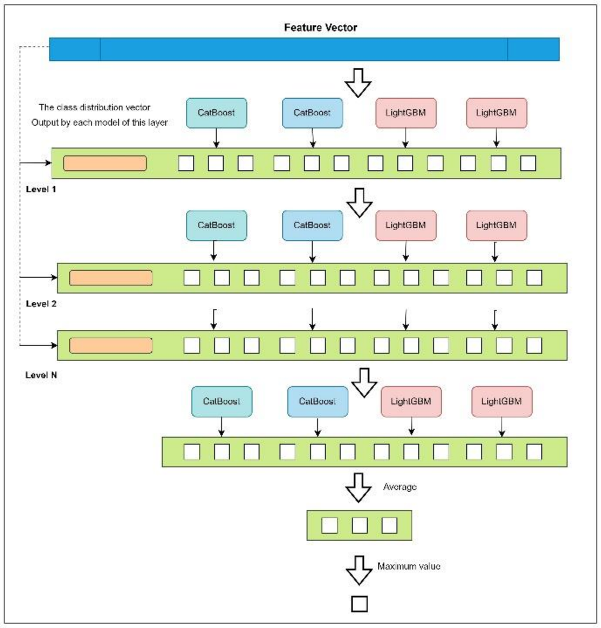

Fig. 1: LHGCAT-XDF algorithm diagram

Influencing Factors

The data derived from reader interactions with e-books on platforms like library portals encompass both basic (personal and borrowing history) and behavioral (search activities and database access) information. This amalgamated data aids in sculpting a comprehensive reader profile, essential for informed e-book purchasing decisions, thereby addressing both explicit and latent reader needs [11] as shown in table 1.

Data Preprocessing

Preprocessing entails organizing the collected data—ranging from e-book details and reader feedback to operational logs and financial records—filtering out pertinent information for dataset creation. The approach involves various technical methods, including web crawlers, to fill in missing values and employs BM25 for correlating words with documents, thus laying the groundwork for the model [12].

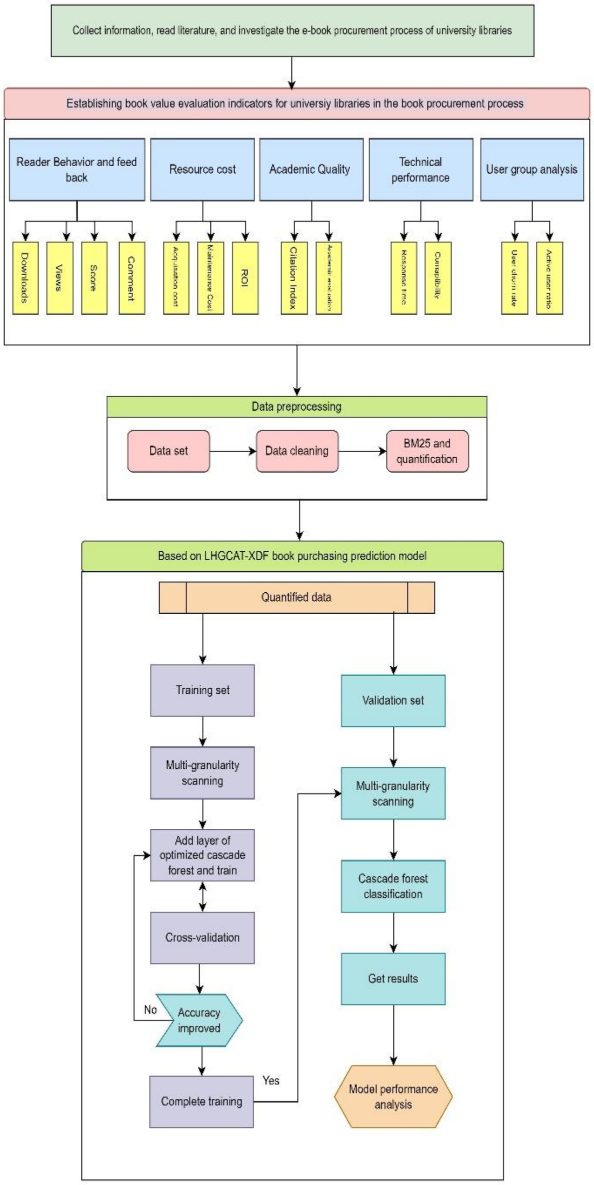
Collect information, read literature, and investigate the e-book procurement process of university libraries
Establishing book value evaluation indicators for universiy libraries in the book procurement process
Reader Behavior and feed back
Resource cost
Academic Quality
Technical performance
User group analysis
Downloads
Views
Score
Comment
Acquisition cost
Maintenance Cost
ROI
Citation Index
Academic evaluation
Response time
Compatibility
User churn rate
Active user ratio
Data preprocessing
Data set
Data cleaning
BM25 and quantification
Based on LHGCAT-XDF book purchasing prediction model
Quantified data
Training set
Validation set
Multi-granularity scanning
Add layer of optimized cascade forest and train
Cross-validation
Accuracy improved
No
Yes
Complete training
Multi-granularity scanning
Cascade forest classification
Get results
Model performance analysis

Figure. 2: Flow chart of e-book purchasing prediction model

Model Prediction and Evaluation

The LHGCAT-XDF-based e-book purchase prediction model operates by initially splitting cleaned data into training and validation sets. Feature selection through multi-granularity scanning precedes training with LightGBM and CatBoost models, incorporating 10-fold cross-validation. The procedure iterates until no significant accuracy improvements are observed. Model performance is assessed using metrics like Accuracy, Precision, Recall, Specificity, and F1 score [13].

Assessment Metrics

When assessing model performance, metrics such as Accuracy, Precision, Recall, Specificity, and F1 score (refer to Table 2 for details) are typically utilized to comprehensively gauge the model's effectiveness

Table 1. The influencing factors in electronic book procurement

Influencing factors	Metrics	Data Sources	Method of Obtaining
user behavior	Downloads Views score Comment	E-book platform backend statistics E-book platform backend statistics e-book platform User review feedback	Download statistics provided by e-book platforms Number of user views recorded through the e-book platform Get ratings through the platform Get feedback from user reviews
resource cost	acquisition cost Maintenance cost ROI	Actual cost of purchasing/subscribing to Maintenance costs of e-book library operations ROI	Obtained from financial records or purchase contracts Obtained from financial records or operating expense details Calculate the benefit and cost ratio of e-book access to obtain
academic quality	citation index academic evaluation	Data from academic databases or citation tools Academic review results	Access through academic databases, citation tools, etc. Obtain from relevant academic publications, journals or platforms
Technical performance	Response time compatibility	E-book platform performance monitoring Platform test report, user feedback	Obtained through performance monitoring tools By conducting platform testing and collecting user feedback
User group analysis	User churn rate Active user ratio	User behavior data analysis tool E-book platform backend statistics	Calculated through user behavior data analysis tools By counting the ratio of the number of active users to the total number of users

Among these, TP (True Positive) represents correctly identified positive instances, TN (True Negative) denotes accurately identified negative instances, FP (False Positive) signifies incorrectly identified positive instances, and FN (False Negative) indicates erroneously identified negative instances [14].

Table 2. Model evaluation criteria

Evaluation Criteria	Meaning	Formula
Accuracy	Represents the ratio of the number of samples that predict the entire sample correctly to the number of the population	TP + TN / TP + TN + FP + FN
Precision	Indicates the proportion of classified positive samples to all samples classified as positive	TP / TP + FP
Recall	Indicates the probability of predicting a positive	TP / TP + FN

	sample among all positive samples	
Specificity	Indicates the proportion of samples correctly predicted as negative class to all actual negative class samples.	TN / TN + FP
F1 value	Expressed as the weighted average of precision and recall	2 x Precision x Recall / Precision + Recall

Experimental Design and Result Analysis

The construction of the deep forest model constitutes a crucial phase, with particular emphasis placed on forest establishment. To enhance model accuracy, various forest parameters require iterative adjustments. The paper introduces LightGBM and XGBoost into the cascade forest structure. LightGBM offers diverse parameter configurations for optimization via cross-validation, where Learning_rate denotes the model's learning rate. A higher learning rate facilitates faster descent along the loss gradient, and vice versa. Num_leaves signifies the number of leaves on each tree, and Max_depth sets the maximum depth of the decision tree regression model. Feature_fraction subsamples features to expedite training and prevent overfitting. Refer to Table 3 for parameter specifications.

Table 3. LightGBM classification model parameter settings

Parameter	Numerical Value	Parameter	Numerical Value
Learning_rate	0.005	Feature_fraction	0.8
N_estimator	927	Num_leaves	10
Max_depth	-1	Max_bin	245
Bagging_fraction	0.6	Bagging_freq	0

By fine-tuning these parameters, the cascade forest was restructured. Multiple experiments were conducted to strike a balance between model runtime and accuracy. Ultimately, the parameter N_estimators = 927 was chosen, resulting in a model accuracy of 79%.

Model Comparison

To underscore the predictive superiority of the deep forest model, traditional learning models (LightGBM, Random Forest, KNN, CNN) were employed to predict sample data, and each model's assessment metrics were compared (refer to Table 4). In terms of specific values, the deep forest achieves an accuracy of 79.0%, markedly surpassing other models. Furthermore, Deep Forest's accuracy of 83.72% also outperforms other models. Recall, specificity, and F1 score also exhibit a notably favorable trend for the deep forest. Although traditional machine learning models boast shorter runtimes, their various assessment metrics pale in comparison to the results obtained with deep forests.

Table 4. Performance evaluation table of various models

Model	LightGBM	random forest	KNN	CNN	LHGCAT-XDF
Accuracy	71.09%	71.5%	69.5%	72.5%	79.70%
Precision	77.63%	72.26 %	76.71 %	77.78 %	83.72%
Recall	0.59	0.61	0.65	0.63	0.90
Specificity	0.83	0.82	0.83	0.82	0.86
F1-Score	67.05%	68.16 %	64.74 %	69.61 %	77.42%

4. CONCLUSION

Accurately predicting electronic book procurement holds paramount importance for university library development. However, existing prediction models suffer from issues of simplicity and low accuracy. To address this, we propose the LIGHT-XDF algorithm, a deep forest algorithm leveraging LightGBM and CatBoost. LightGBM and CatBoost are integrated into the cascade forest, with CatBoost enhancing prediction accuracy and LightGBM reducing model complexity. The LIGHT-XDF algorithm utilizes reader behavioral data and electronic library collection data for purchase predictions. Experimental findings demonstrate that, compared to alternative models, LIGHT-XDF exhibits superior overall performance. Future endeavors will focus on validating the robustness and generalization capability of the LIGHT-XDF algorithm through extensive performance testing on diverse library collection datasets. Additionally, exploration of various new technologies will be pursued to enhance the accuracy of overall e-book procurement predictions.

References

[1] Li, M. (2007). Application of web-based data mining technology in digital libraries. Journal of Academic Library and Information Science, 25, 44-46.

[2] Affum, M. Q. (2023). Book Acquisition in the Modern University Library: Challenges and Opportunities. Library Philosophy & Practice.

[3] Anna, N. E. V., & Mannan, E. F. (2020). Big data adoption in academic libraries: a literature review. Library Hi Tech News, 37(4), 1-5.

[4] Zhou, Z. H., & Feng, J. (2019). Deep forest. National science review, 6(1), 74-86.

[5] Ma, W., Yang, H., Wu, Y., Xiong, Y., Hu, T., Jiao, L., & Hou, B. (2019). Change detection based on multi-grained cascade forest and multi-scale fusion for SAR images. Remote Sensing, 11(2), 142.

[6] Hancock, J. T., & Khoshgoftaar, T. M. (2020). CatBoost for big data: an interdisciplinary review. Journal of big data, 7(1), 94.

[7] Liang, Y., Wu, J., Wang, W., Cao, Y., Zhong, B., Chen, Z., & Li, Z. (2019, August). Product marketing prediction based on XGboost and LightGBM algorithm. In Proceedings of the 2nd international conference on artificial intelligence and pattern recognition (pp. 150-153).

[8] Jabeur, S. B., Gharib, C., Mefteh-Wali, S., & Arfi, W. B. (2021). CatBoost model and artificial intelligence techniques for corporate failure prediction. Technological Forecasting and Social Change, 166, 120658.

[9] Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. Foundations and Trends® in Information Retrieval, 3(4), 333-389.

[10] Zhao, F., Kumamoto, E., & Yin, C. (2021, July). The effect and contribution of e-book logs to model creation for predicting students' academic performance. In 2021 International Conference on Advanced Learning Technologies (ICALT) (pp. 187-189). IEEE.

[11] Trakarnsakdikul, N., Chaiyaphan, S., & Leecharoen, B. (2021). Factors Affecting E-Book Purchase Decisions of Customers in Thailand. Asian Administration & Management Review, 4(1).

[12] Whissell, J. S., & Clarke, C. L. (2011). Improving document clustering using Okapi BM25 feature weighting. Information retrieval, 14, 466-487.

[13] Torgo, L., & Ribeiro, R. (2009). Precision and recall for regression. In Discovery Science: 12th International Conference, DS 2009, Porto, Portugal, October 3-5, 2009 12 (pp. 332-346). Springer Berlin Heidelberg.

[14] Vujović, Ž. (2021). Classification model evaluation metrics. International Journal of Advanced Computer Science and Applications, 12(6), 599-606.