Research on the Application of Artificial Intelligence Based Technology in Chinese Book Procurement in Universities—Take Shaoyang University Library as an Example

Ke Luo(✉)

Department of Library, Shaoyang University, Shaoyang, Hunan, China
luoke00@qq.com

Abstract. To reduce the risk of zero borrowing of purchased books, the live use of library space, and the optimization of the library's shelving procedures, given the limited funds and space in university libraries, this article uses algorithms such as Light Gradient Boosting Machine (LightGBM), Support Vector Machine (SVM), and K-Nearest Neighbor (KNN) for classification and regression The analysis was conducted and the results were cross-checked 10 times. By comparing the experimental results, the LightGBM method has an accuracy of 73.17% and a RMSE value of 4.154, and the model can be used as a basis for effective purchasing decisions.

Keywords: Artificial intelligence technology · university book procurement · machine learning

1 Introduction

With the deepening of the construction of university intelligent libraries, library resources gradually shift from collection storage space to learning space gradually, promoting the balanced transfer of paper and electronic resources [6], making the paper collection must be used more effectively in the space occupied by the library. High school libraries generally have limited storage space for collections and insufficient funds, and electronicization and precise purchase of paper resources have become two ways to solve the problem [2].

At present, most of the university libraries in China are still in the "subjective-led, technology-assisted" interviewing mode, in which the titles of book purchases are determined according to the annual book acquisition funds, the experience of interviewers, the suggestions of teachers and students, and the recommendations of merchants [9, 11]. In this traditional model, it is still mainly dependent on the subjective will and experience of the book interviewer to decide the number of books to be purchased and the specific titles, and modern information technology has not been fully integrated into the interviewing business.

Z. Zhan et al. (Eds.): ICBDIE 2022, AHCS 5, pp. 936–946, 2023.
https://doi.org/10.2991/978-94-6463-034-3_96

Indeed, various information technologies have become increasingly mature nowadays, especially the development and popularity of AI-assisted decision making, which has rewritten the traditional operating models and ecosystems of various industries [9]. Likewise, it provides new ideas for the innovation of book acquisition mode and management of university libraries and clarifies the future construction direction of university libraries [1]. Future library services should be able to understand not only the collections but also the readers better, rather than just catering to people with different preferences, and understanding their needs will become an important part of library development. Machine learning, as the core and hottest technology in the field, can automatically identify patterns and discover rules based on a large amount of data, predict readers' willingness to borrow, and provide the possibility to meet readers' needs for personalized learning. In this paper, BM25 model corresponding to textual information is used for feature engineering, and algorithms such as Light Gradient Boosting Machine (LightGBM), Support Vector Machine (SVM), and K-Nearest Neighbor (KNN) are applied for classification and regression analysis to build a system of accurate purchase prediction models based on the evaluation results. The design concept is that by inputting new recommended book list information, the system obtains the best purchased titles based on the prediction results and set thresholds, which can assist librarians to process the recommended book lists more quickly, reduce the risk of books not being checked out, and maximize the accuracy of predicting book checkouts in order to reduce the purchase of duplicate paper resources and improve the overall checkout rate of the library.

2 Related Work

2.1 Light Gradient Boosting Machine

Gradient boosting decision tree (GBDT) is one of the most widely used methods in current machine learning, and Light Gradient Boosting Machine (LightGBM) is a new technique developed in machine learning in recent years, which was originally introduced by GuolinKe and his research group at Microsoft Research in 2017 as a LightGBM differs from the original gradient boosting judgment tree (GBDT) [5] in that it is highly distributed and efficient for classification, regression, and ranking of large-scale, multidimensional data. The basic principles of LightGBM include the following two particular techniques.

Gradient-based One-Side Sampling (GOSS), which randomly rejects most samples with small gradients by setting a threshold, because the size of the gradient has a direct effect on the calculation of information gain, GOSS keeps those samples with large gradients (e.g., greater than a predetermined threshold, or between the highest percentile) and randomly removes those samples with small gradients. This results in more accurate gain estimates than uniform random sampling, with the same target sampling rate, especially when the range of values of information gain is large.

Exclusive Feature Bundling (EFB) is a method of feature compression in a sparse feature space, where generally high-dimensional data are sparse, by combining two mutually exclusive features into a new one, thus reducing the learning speed and complexity of the algorithm. When two features cannot be completely exclusive, a conflict rate is computed to measure the degree of feature non-repulsion.

LightGBM differs from the level-wise algorithm of GBDT in its decision tree generation algorithm and uses the depth-limited leaf-wise algorithm. Level-wise looks for the leaf with the largest splitting gain at that level to grow. Compared with level-wise, leaf-wise has the advantage of significantly higher accuracy, while the disadvantage is that it may produce very deep decision trees leading to overfitting, so LightGBM uses additional independent variables in the leaf-wise algorithm to limit the depth and avoid overfitting.

2.2 BM25

BM25, also known as Okapi BM25, is based on the probabilistic retrieval framework developed by Stephen Robertson and Karen Spärck Jones et al. in the 1970s and 1980s. The BM25 [10] algorithm is applicable to calculate the correlation between a certain target text file and unlike TF-IDF, BM25 consists of three main components: the weight of a word, the relevance between the word and the target document, and the relevance between the word and the query keyword. The product of these three components forms the score of a word, and the sum of the scores of all words in the query keyword becomes the relevance score of the whole document with respect to a query keyword. The algorithm of BM25 is formulated as follows.

$$Score(D, Q) = \sum_{i=1}^{n} IDF(q_i) \times \frac{f(q_i, D) \times (k_1 + 1)}{f(q_i, D) + K_1 \times (1 - b + b \times \frac{L_d}{L_{avg}})} \tag{1}$$

where $f(q_i, D)$ is the word frequency of word q_i in document D, L_d is the length of document d, L_{avg} is the average length of all documents, and the variable K_1 is a positive parameter, usually k_1 will be set between 1.2 and 2.0, and b = 0.75. $IDF(q_i)$ is then the word weight, denoted as follows.

$$IDF(q_i) = \ln(\frac{N - n(q_i) + 0.5}{n(q_i) + 0.5} + 1) \tag{2}$$

Where N denotes the number of all documents in the index, $n(q_i)$ is the number of documents contained, and the numerator denominator is added with 0.5 for smoothing. According to the role of IDF, for a q_i, the more documents containing q_i, the less important q_i, or the lower the distinction, the smaller the IDF, so IDF can be used to characterize the similarity of q_i and documents.

3 Methodology

3.1 Building an Accurate Book Purchasing Model Based on Machine Learning Algorithms

In this paper, the borrowing records of Shaoyang College Library for ten years (including 13 fields such as bibliographic system number, book title, author, publisher, year of publication, and total number of borrowings) are used as the research object to construct an accurate acquisition model. We will use BM25 to feature engineer the text data, and

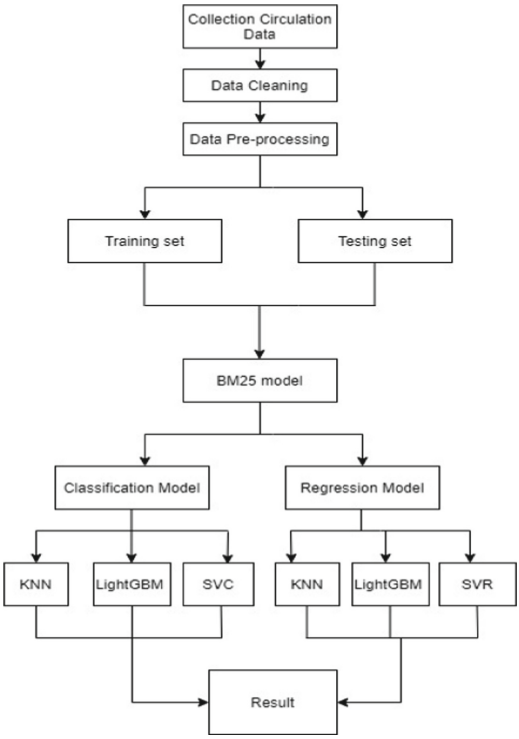

Fig. 1. Research Flow Chart

then construct regression and classification models with LightGBM, KNN and SVM methods, and the model design process is shown in Fig. 1.

Step 1: Data cleaning of the original data set. Because of the large amount of original data and book data processing problems, some fields have missing values; if there are missing values, they are directly deleted.

Step 2: Pre-processing the cleaned data with data such as punctuation and stop word cleaning.

Step 3: dividing the cleaned data set into a training set and a test set.

Step 4: Combine the four fields of book title, author, publication year and publisher in the training and test sets, and use jieba to break the Chinese words.

Step 5: Obtain the correlation between words and documents using BM25.

Step 6: Construct a model based on the correlations obtained in Step 5 using LightGBM, KNN and SVM methods, and perform 10 times of cross-validation to obtain model accuracy and evaluate various pointers such as regression models.

Step 7: Evaluate the model performance based on the prediction results.

3.2 Model Evaluation Standards

3.2.1 Confusion Matrix

The confusion matrix [8] is an error matrix, mainly by looking at the predicted and correct categories, with a table stating the number or percentage of correct and incorrect predictions for each category, and is commonly used in binary or multivariate classification problems. As shown in Table 1.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

3.2.2 Classification Model Metrics

Many metrics can be derived from the confusion matrix to calculate the precision, recall, and F1 scores to evaluate the merits of this model. Precision is the number of correct classifications as a percentage of all the numbers classified as correct, as shown in Eq. (4); recall is the chance of how many positive samples are predicted among all positive samples [7], as shown in Eq. (5); and F1 scores is a comprehensive evaluation metric [3], combining precision and recall, as shown in Eq. (6).

$$\text{Precision Rate} : P = \frac{TP}{TP + FP} \tag{4}$$

$$\text{Recall Rate} : R = \frac{TP}{TP + FN} \tag{5}$$

$$\text{F1 scores} : F1 = \frac{2PP}{P + R} \tag{6}$$

3.2.3 ROC and AUC

The ROC curve is mostly used for binary classification and is able to evaluate multivariate classification, reflecting the relationship between false positive rate and true positive rate. Its X coordinate axis indicates the false positive rate and Y coordinate axis indicates the true positive rate. The closer the ROC curve is to the y coordinate axis, the better the prediction performance of the model. The area under curve (AUC) is usually used to measure the prediction accuracy of the model, and the larger the value of AUC, the higher the correct rate.

Table 1. Confusion matrix

Confusion matrix		Predicted Class	
		Positive	Negative
Actual Class	Positive	TP	FN
	Negative	FP	TN

Table 2. Operating environment

Project	Version
Operating System	Mircosoft Windows 10
Processor	Inter® Pentium® CPU G3250 @3.20 GHz
Storage Space	1 TB
Database	Oracle
Development Language	Python 3.8 version

4 Experiment and Performance Evaluation

4.1 Experimental Environment

The experimental environment is conducted under Windows 10 operating system, and the database is used on Oracle for data access and query, and Python version 3.8 is used for text content processing and model construction and other related research work. As shown in Table 2.

4.2 Experimental Design and Analysis of Results

The experiments in this paper consist of three main components, firstly, data set selection, secondly, description of data cleaning rules, and finally, the relevance scores for each keyword were calculated using BM25 as the independent variables of LightGBM, KNN and SVM models, and the performance of both classification and regression models was evaluated by parameter optimization and 10 times cross-validation.

4.2.1 Data Set Selection

The text data set used contains 13 fields such as bibliographic system number, book title, author, year of publication, publisher, classification number, price, total number of borrowings, etc. According to the total number of borrowings, a new field is added as a classification, and no borrowed books are defined as 0, while the classification of books with borrowings is defined as 1. There are 103948 items through the collation. 80% of the original data are classified as the training set and the rest as the test set.

4.2.2 Data Cleaning

The collated book circulation data will be cleaned as follows:

(1) Missing values: in the process of data collection, especially in the case of a large amount of data, there is a high possibility of missing data, screening text data, if the book field is blank, it should be deleted.
(2) Non-book data: In the database lending records, there will be lending records such as CD-ROMs, which are not considered in this study and should be deleted.

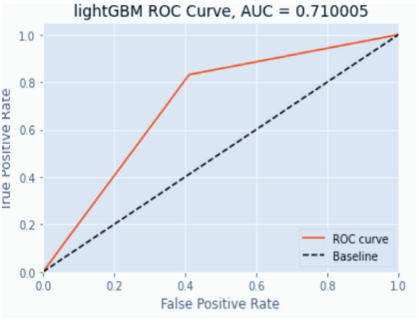

Fig. 2. LightGBM ROC Curves

(3) Deactivated words: Deactivated words refer to words that occur frequently but are meaningless, such as: of, had, in, is, have, and so on, retaining the key words, and the rest of the deactivated words should be deleted.
(4) Non-text data: data fields with a lot of non-text data, such as punctuation, accent marks, blank areas and non-English characters, should be deleted.
(5) Data normalization: data normalization refers to scaling the text data and setting the value range between [0, 1] to reduce the training time and improve the accuracy of the model.

4.2.3 Model Performance Evaluation

4.2.3.1 Classification Model Evaluation

Model 1 is LightGBM: LightGBM is a machine learning gradient boosting model with better efficiency and higher accuracy, suitable for processing huge amount of 7 data. Fraction is the fraction of features needed for each iteration set to 0.8, bagging_fraction is the data needed for each iteration to improve the training speed of the model and avoid overfitting set to 0.6, max_bin is the maximum amount of loaded values set to 245, and the other parameters are listed in Table 3. The confusion matrix and ROC of the model are shown in Fig. 2, and the average accuracy is 73.17%.

Model 2 is KNN: The central idea of the K nearest neighbor algorithm is that a text data to be classified finds K most similar neighbors in the text feature space, and most of the neighbors are classified into a certain class, then the text data to be classified is of that class. By cross-validation, the model parameter k value is set to 9, and the average accuracy is obtained as 55.68%, and the ROC graph of KNN is shown in Fig. 3.

Model 3 is SVC, which is a supervised learning method used by SVM to deal with classification problems, and SVM is a supervised learning method that finds a decision boundary on the hyperplane to maximize the boundary between two classifications and make them perfectly distinguishable. The parameter C is the penalty coefficient of the model, and the larger the coefficient, the higher the penalty for classification errors, and the higher the accuracy in post-training tests, but it is prone to overfitting. In this study, the model parameter C is set to 1.0, and the average accuracy of the model is 66.1% by cross-validation and SVC confusion matrix, as shown in Fig. 4 and Table 4.

Table 3. LightGBM classification model parameter

Parameter	Numeric	Parameter	Numeric
max_depth	7	verbose	-1
min_data_in_leaf	11	bagging_freq	0
feature_fraction	0.8	cat_smooth	0
bagging_fraction	0.6	min_split_gain	0.6
lambda_l1	0.5	n_estimators	927
lambda_l2	0.001	max_bin	245
objective	binary	metric	auc
boosting	gbdt	num_leaves	10
num_boost_round	1000	learning_rate	0.1

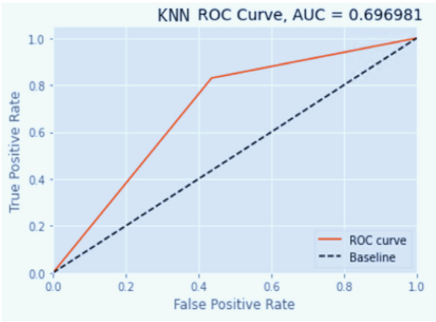

Fig. 3. KNN ROC Curves

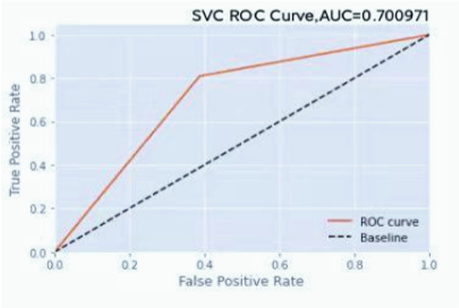

Fig. 4. SVC ROC Curves

Table 4. Classification model performance evaluation

Model	LightGBM	SVC	KNN
Accuracy	73.17%	66.1%	55.6%
Precision	0.712	0.642	0.619
Recall	0.706	0.601	0.551
F1 scores	0.763	0.574	0.515

Table 5. LightGBM regression model parameter

Parameter	Numeric	Parameter	Numeric
Learning_rate	0.05	Feature_fraction	0.8
N_estimator	927	Num_leaves	50
Max_depth	8	Bagging_freq	5
Bagging_fraction	1.0		

4.2.3.2 Regression Model Evaluation

The borrowing data were then subjected to regression analysis to predict the likely future borrowing of the books in the collection, again using LightGBM, SVR, and KNN, respectively (Table 5).

The mean square error (MSE) is 17.252, the root mean square error (RMSE) is 4.154, and the mean absolute error (MAE) is 1.712 for the model trained by the above parameters.

Model 2 uses the SVR model, which is also an important application branch of SVM, to extend the regression method to a hyperplane and then calculate the total distance between each data point and the hyperplane, with a parameter setting of C of 10. The mean square error (MSE) is 19.651, the root mean square error (RMSE) is 4.427, and the mean absolute error (MAE) is 1.776 after the calculation of the predicted and correct values of the model. 1.776.

Model 3 uses KNN. In addition to solving classification problems, KNN can also be used for regression problems, also based on the nearest data point as a prediction. The mean square error (MSE) is 17.517, the root mean square error (RMSE) is 4.193, and the mean absolute error (MAE) is 1.763 for the parameter optimization case with the K value set to 25 (Table 6).

Table 6. Regression model performance assessment

Model	LightGBM	SVR	KNN
MSE	17.252	19.651	17.517
RMSE	4.154	4.427	4.193
MAE	1.712	1.776	1.763

5 Conclusion

In this paper, three machine learning methods: LightGBM, SVM and KNN were used to predict the classification and regression results of the number of borrowed books by using the Chinese collection borrowing data provided by Shaoyang College Library for the past ten years. In the classification problem, LightGBM has the best accuracy rate of 73.17% among the three machine learning methods, which provides managers with aids for purchasing Chinese books. In the regression analysis, LightGBM also had the best results, with a root mean square error (RMSE) of 4.154 and a mean absolute error (MAE) of 1.712. The machine learning regression method predicts the number of possible checkouts, allowing librarians to be more effective in handling the book shelving process and enabling patrons to read the books they want faster.

This paper analyzes and predicts the number of borrowings from the paperback collection in order to avoid the library limiting the patronage to current students, which makes the model over-fitted and generates decision errors after the students' enrollment and graduation rotation. In the future, the model can be updated with the latest trendy keywords and more diverse book categories by adding reference factors such as the number of purchases, browsing and recommended book rankings on online book purchasing or reading promotion platform websites.

Acknowledgment. This paper is funded by Hunan Provincial Philosophy and Social Science fund project "Research on the Application of Artificial Intelligence Technology in Accurate Book Procurement in Universities under Smart Library" (21YBA179).

References

1. Bai Guangsi(2016). Research on the library intelligent resource procurement systerm based on big data. Research on Library Science, 19,37-41.
2. Buckland, M. K.(2017). Library technology in the next 20 years. Library Hi Tech,35(1),5–10.
3. Chicco, D., Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics, 21(1),1–13.
4. Duan, Y., Edwards, J. S., Dwivedi, Y. K.(2019). Artificial intelligence for decision making in the era of Big Data -evolution, challenges and research agenda. International Journal of Information Management, 48,63-71.
5. G. Ke, Q. Meng, T. Finley, T.Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu, 2017. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. 31st Conference on Neural Information Processing Systems, NIPS 2017 ,3146–3154.

6. Latimer K(2010). Redefining the library: current trends in library design. Art Libraries Journal ,35(1), 28–34.
7. Liu, Z., Bondell, H. D. (2019). Binormal Precision–Recall Curves for Optimal Classification of Imbalanced Data. Statistics in Biosciences: Journal of the International Chinese Statistical Association, 11(1),141-161.
8. Mi Aizhong, Zhang Pan(2017). A method of classifier selection based on confusion matric.Journal of Henan Polytechnic University(Natural Science) ,36(2) ,116–121.
9. Runhua Wang,Yi Tang,Lei Li(2012). Application of BP neural network to prediction of library circulation.Cognitive Informatics & Cognitive Computing ,117(3),31–39.
10. Stephen Robertson, Hugo Zaragoza(2009). The Probabilistic Relevance Framework:BM25 and Beyond. Foundations and Trends in Information Retrieval.3 (4),333–389.
11. Susan Thompson(2012). Student Use of Library Computers: Are Desktop Computers Still Relevant in Today's Libraries? Information Technology and Libraries, 31(4),20–33.