Financial Budget Item Identification Model: Accurately Matching the Budget Items of Reimbursement Claims based on KNN Algorithm

Junyi Wáng[1,a], Peichun Suo[2,b*], Weili Kou[1,c*],Yan Zhang[3,d], Meicai Zhu[4,e]

[1]College of Big Data and Intelligent Engineering, Southwest Forestry University, Kunming, Yunnan, China
[2]Information Center, Yunnan Medical Health Vocational College, Kunming, Yunnan, China
[3]The 11th Middle School of Qilin District Qujing City China
[4]Student Affairs Office, Yunnan Medical Health Vocational College, Kunming, Yunnan, China

[a]25807424@qq.com; [b*]suopeichun_ynzt@163.com;
[c*]kwl@swfu.edu.cn; [d]405917503@qq.com; [e]3344328962@qq.com

Abstract. This study aimed to scrutinize the procedural mechanisms underlying the accurate selection of financial budget items in reimbursement forms, to reinforce adherence to predefined expenditure guidelines, enhance the standardization of form completion procedures, and augment work efficiency. The 7,959 reimbursement dataset was collected from a medical vocational college in Yunnan Province of China in 2022 and partitioned into training (80%) and testing (20%). Leveraging Random Forest (RF) complemented by Recursive Feature Elimination (RFE), we selected pertinent attributes. Further, utilizing Grid Search CV, we trained a K-Nearest Neighbors (KNN) predictive model for budget item classification and contrasted its performance with a Decision Tree (DT) and GBDT. Our outcomes illuminated that the prediction accuracy of KNN (81%) is higher than both GBDT (42%) and DT (59%). These insights offer fresh perspectives on the procedural dynamics of financial budgeting, potentially informing strategies for improving financial management practices and form processing efficiency.

Keywords: Budget Items, Grid Search CV, KNN, Decision Tree, GBDT

1 Introduction

In modern times, the financial budgeting system stands as a pivotal nexus bridging an organization's strategic aspirations with its daily operational fabric. Strengthening the governance of financial budget projects emerges as a strategic imperative for enhancing organizational stewardship, fostering efficient resource allocation, and upholding fiscal well-being.

Simultaneously, the realm of machine learning has infiltrated myriad disciplines, reshaping traditional paradigms, including those in finance.[1]AlHajri showcased machine learning's versatility in indoor localization.[2]Lussier highlighted its analytical

© The Author(s) 2024

E. P. H. Lau et al. (eds.), *Proceedings of the 2024 3rd International Conference on Information Economy, Data Modelling and Cloud Computing (ICIDC 2024)*, Advances in Computer Science Research 114,
https://doi.org/10.2991/978-94-6463-504-1_10

expertise in complex spectroscopic data.[3]Liu delved into its potential in battery material design.[4]Huang explored its application in construction engineering. [5]Yang its role in IoT networks.[6]Demrozi in human activity recognition.[7]Hu demonstrated its practicality in financial forecasting and Searle stressed the need for privacy-ensured federated learning in sensitive sectors among these advancements, K-Nearest Neighbors (KNN) algorithms have carved a niche in financial analytics, with scholars refining methods for distress prediction, stock market analysis, and financial forecasting. Highlighting KNN's adaptability and effectiveness, these studies underscore its value in financial decision-making.

Building upon this foundation, our investigation employs 2022 financial reimbursement text data from a private Yunnan university, employing an 8:2 data split. We construct a KNN-based classification model, validated through 5-fold cross-validation, with the ambition to the reimbursement data deeply, establish a KNN model tailored for higher education’s financial reimbursement, and ultimately evaluate it to yield accuracy indices of 79% in training and 80.65% in testing. This not only protects the accuracy of financial data analysis but also furnishes a practical blueprint for enhancing financial management practices within academic institutions.

2 Research Methods

2.1 K-Nearest Neighbors (KNN)

The mathematical cornerstone of the algorithm lies in an effective distance metric, and the common metrics are Euclidean distance, Manhattan distance, and the broader Minkowski distance, among which Euclidean distance is the preferred solution due to its intuition and wide applicability, whose mathematical expression can be referenced in Eq. (1).

$$d(x_1,x_2)=\sqrt{\sum_{j=1}^{n}\left(x_{1j}-x_{2j}\right)^2} \tag{1}$$

Let n denote the number of feature dimensions. Suppose x_l and x_2 represent the values of two samples on the j feature dimension. To quantify their dissimilarity on this dimension, we calculate the square of the difference between these values. Proceeding through all n features, we accumulate the squares of differences across dimensions and subsequently take the square root of this aggregate, yielding the Euclidean distance between the two sample points. A smaller distance signifies closer proximity between the two samples within the feature space.

2.2 Datasets

This study analyzed 7,959 reimbursement records from a medical vocational college in Yunnan in 2022, involving 15 approval processes and over 100 secondary budget items. Reimbursement records from 14 departments, grouped under main budget projects, made up the data. The dataset initially comprised 80 columns. Following preprocessing, empty columns were removed, narrowing down the dataset to 30 columns, consisting of 22 categorical variables, 6 numerical variables, and 2 date-time types. Lastly, the dataset was streamlined to retain 14 informative columns, namely abstract, account name, handler, reversal indication, sub-project name, main project name, budget item name, document creator, creation timestamp, reviewer, review timestamp, major project category, supervisor name, and department affiliation. Employing an 8:2 ratio, the dataset was partitioned into subsets designated for training and testing purposes, with the training set encompassing 6,367 samples, representing 80% of the total, and the test set containing 1,592 samples, accounting for the remaining 20%. These subsets spanned diverse textual data about various reimbursement items.

2.3 Feature Selection

We are implementing the feature selection process as shown in Fig.1. Firstly, CSV data files containing Chinese characters are read through the Pandas library, and Chinese word splitting is processed using Jieba. Then, the TF-IDF feature matrix is constructed for each processed text field using TfidfVectorizer. Afterward, the dataset is partitioned into training and test sets, and the key features are screened out using a Random Forest Classifier with Recursive Feature Elimination (RFE), which is a method that gradually removes unimportant features based on the model evaluation metrics, and the most important 14 features are set to be selected here. Based on the features filtered by RFE, we further evaluate them using a Random Forest Classifier. Finally, the Seaborn library is used to generate a bar chart showing the importance scores of the features selected by RFE in the RF model as shown in Fig.2, which helps to understand the degree of contribution of each feature to the "main item name". Finally, five features (Summary, Master Item Name, Handler, Project Leader, Department) were identified for selection in conjunction with the Reimbursement Form Fill Item.

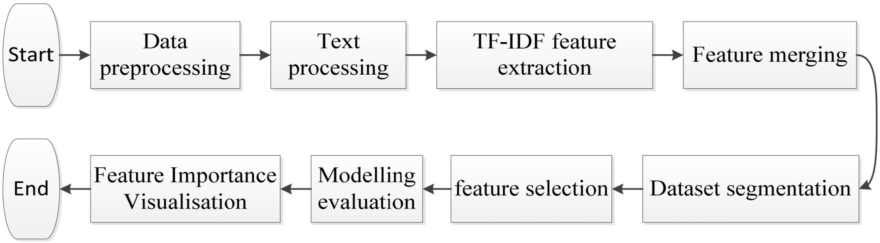

Fig. 1. Feature Selection

2.4 Model Construction

The study used K-Nearest Neighbors (KNN) to construct the model to match financial reimbursement expenditures with corresponding budget items. The model development encompassed four pivotal stages: data preprocessing, feature extraction, model establishment, and performance assessment (Fig.3). Firstly, we load the local CSV file and segment the textual fields to comprising the main project name, project abstract, handler, responsible party, and department. Secondly, we transform these segmented text fields into numerical feature vectors by the TF-IDF vectorization technique. Thirdly, KNN was used to conduct classifications. Thirdly, we verified both training and testing results by a 5-fold cross-validation strategy (k=5). The verification metrics include accuracy, precision, recall, and the F1 score.

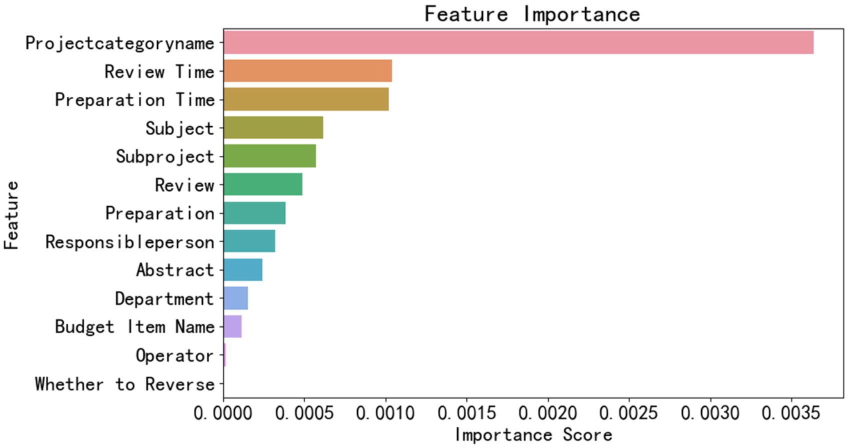

Fig. 2. Feature Importance Ranking

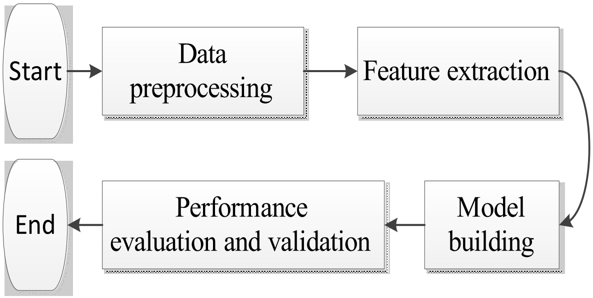

Fig. 3. Model construction

2.5 Model Parameter Optimization

We commence by constructing a grid of parameters, stipulating a range for K (n_neighbors) to encompass all integer values from 1 to 30. Leaning on K-fold cross-validation, the model undergoes repetitive training and evaluation across varied parameter configurations, a strategy designed to mitigate randomness and bolster reliability in assessments. Ultimately, Grid Search CV identifies the parameter constellation that excels across these evaluative criteria.

2.6 Model Evaluation

The predictive performance and accuracy of the model in this study are quantitatively assessed through four key metrics: Accuracy (Eq.2), Precision (Eq.3), Recall (Eq.3), and the F1 Score (Eq.4).

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN}) \times 100\% \quad (2)$$

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP}) \times 100\% \quad (3)$$

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN}) \times 100\% \quad (4)$$

$$\text{F1} = (2 * \text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall}) \times 100\% \quad (5)$$

where, TP, FP, FN, and TN denote the counts of true positives, false positives, false negatives, and true negatives in the predictions, respectively, forming the fundamental components for evaluating classification model performance.

A 5-fold cross-validation was employed for model verification, while both the cross-validation and Grid Search CV were leveraged to optimize the parameters of the KNN model. Ultimately, the performance of the KNN model was benchmarked against the Decision Tree and Gradient Boosting Decision Tree (GBDT) (Table 1).

3 Results and Analysis

3.1 Model Evaluation

We use the KNN model to predict 1592 real claim application test set data, and finally 1283 correct predictions and 309 incorrect predictions, with an accuracy rate of 80.5904% and an error rate of 19.4095%.

According to feedback from the Treasury, each telephone inquiry on a claim takes about 60 seconds, and under normal circumstances, there are about 60 claims per day, which would require one hour per person-day if telephone inquiries were used. According to the calculation of 8 hours of work per person day, of which 1 person per day in addition to answering the phone, 7 hours of normal work; assuming that each person's monthly salary of 5,000 thousand, 22 days of normal working days per month, 8 hours of work per day, the average hourly wage on a normal working day of 28.4 yuan, 1 person per month in the telephone consulting of the required salary of about 625; the use of machine learning KNN model for the booking of appointments after the error by the Prediction 19.4095% calculation, 1 person per month in the telephone consultation of the required salary of about 121.30 yuan, 1 person per month in addition to saving the salary expenditure of 503.7 yuan, the number of telephone consultation is reduced by nearly 8 times. At the same time, reducing telephone communication and avoiding the impact of human negative emotions brought about is conducive to improving the quality of work and life and the impact of the physical and mental health of faculty members.

3.2 Model Comparisons

In this study, focusing on the prediction classification task of university financial reimbursement projects, we evaluated and compared three machine learning models – KNN, DT, and GBDT on the test set using four key performance metrics: Accuracy, Precision, Recall, and F1-score, the results are shown in Table 1. The results indicated that the KNN algorithm demonstrated superior accuracy compared to the Decision Tree and GBDT models in classifying the financial budget items.

Table 1. Model comparison

Model	Overall Accuracy	Precision	Recall	F1 Score
KNN	0.81	0.86	0.81	0.79
GBDT	0.42	0.91	0.42	0.37
DT	0.59	0.95	0.59	0.59

4 Conclusions

This study collected 7,959 financial reimbursement records from a medical vocational college of Yunnan in 2022, identified five crucial features, and employed the KNN to construct a budget classification model, achieving a testing accuracy of 81%, surpassing Gradient Boosting Decision Tree's (GBDT) 42% and traditional Decision Tree's 59%. The KNN model also excelled in precision, recall, and F1-score, highlighting its classification expertise. The research provides a model that informs financial process optimization, ensures compliance, and enhances efficiency, thereby serving as a reference for educational and broader financial management sectors. Future endeavors look towards integrating deep learning models to further enhance classification accuracy and generalize performance.

Acknowledgment

This research was funded by the Scientific Research Fund of Yunnan Provincial Department of Education under Research and Development Project grant number 2024J2141.

References

1. M. I. AlHajri, N. T. Ali, and R. M. Shubair, "Indoor Localization for IoT Using Adaptive Feature Selection: A Cascaded Machine Learning Approach," IEEE Antennas and Wireless Propagation Letters, vol. 18, no. 11, pp. 2306-2310, 2019. https://doi.org/10.1109/LAWP.2019.2915047.

2. F. Lussier, V. Thibault, B. Charron, G. Q. Wallace, and J.-F. Masson, "Deep learning and artificial intelligence methods for Raman and surface-enhanced Raman scattering," Trends in Analytical Chemistry, vol. 124, pp. 115796-115796, 2020. https://doi.org/10.1016/j.trac.2019.115796.
3. Yue Liu, Biru Guo, Xinxin Zou, "Machine learning assisted materials design and discovery for rechargeable batteries," Energy Storage Materials, 2020, 31:434-450. https://doi.org/10.1016/j.ensm.2020.06.033.
4. Huang M.Q. et al, "BIM, machine learning and computer vision techniques in underground construction: Current status and future perspectives," Tunnelling and Underground Space Technology 108(2021):103677-. https://doi.org/10.1016/j.tust.2020.103677.
5. Yang Kai, Shi Yuanming, Zhou Yong, "Federated Machine Learning for Intelligent IoT via Reconfigurable Intelligent Surface," IEEE NETWORK, 2020, 34(5):16-22. https://doi.org/10.48550/arXiv.2004.05843.
6. Demrozi Florenc, Pravadelli Graziano, Bihorac Azra, et al. Human Activity Recognition using Inertial, Physiological and Environmental Sensors: A Comprehensive Survey.[J]. IEEE access: practical innovations, open solutions, 2020, 8:210816-210836. https://doi.org/10.1109/ACCESS.2020.3037715.
7. Hu Zhengbing, Uhryn Dmytro, Ushenko Yurii, et al. Corporate information system for exchange rate analysis and commodity money forecasting. 2023, 12938:129380N-129380N-4. https://doi.org/10.1117/12.3009679.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

BY NC