Random Forest Algorithm for HR Data Classification and Performance Analysis in Cloud Environments

Fangfang Dong
College of Management, Zhengzhou Shengda University of Economics,
Business and Management, Zhengzhou 451191, Henan, China

***Abstract*—This study applies the Random forest algorithm to classify and evaluate the effectiveness of business human resources (HR) data, focusing on its potential in supporting strategic decision-making and enhancing organizational efficiency. The research introduces a model that automates the categorization of HR data, including employee records, performance evaluations, and training activities, using the Random Forest method. By constructing both classification and effectiveness assessment models, the study aims to provide businesses with a robust tool for managing and evaluating employee contributions. Key HR metrics were analyzed and categorized, leading to the creation of an effectiveness evaluation model that offers objective insights into employee performance. The Random forest algorithm's accuracy and stability were validated through cross-validation techniques, proving it to be effective in categorizing employee data and identifying different workforce groups. The models developed in this study are designed to support HR managers in optimizing human resource allocation, improving employee satisfaction, and driving overall business performance. The paper also discusses how the model can be optimized further by expanding data sources and applying it to practical business scenarios.**

Keywords—Random forest algorithm; business; human resources; data classification

I. Introduction

In recent years, the emphasis on investment and R&D programs has become more pronounced as China moves from a rapid growth model to one that pursues high-quality development. Between 2013 and 2015, the formalization of China's R&D workforce increased rapidly. The growth of R&D companies and the rapidly changing needs of users have contributed significantly to the increase in R&D projects. In order to compress the R&D cycle, it has become an emerging trend in the industry for multiple R&D companies to develop multiple projects in parallel at the same time [1]. For many projects, each project's overall status and profitability must be considered thoroughly. Therefore, achieving high performance with limited resources is much more complex than in a single-design environment and requires more than a rational layout [2]. The planning of human resources of R&D personnel as the main body of innovation takes priority in electronic projects, which highly depends on the rational allocation of resources. However, project planning and human resource allocation have become more complicated due to multitasking conflicts and the shortage of R&D personnel [3]. Construction delays and budget overruns are often triggered when the schedule and staffing cannot be synchronized. Therefore, how to rationally organize the task planning and staff allocation of multiple projects, complete the overall construction cycle in the shortest time, improve the company's economic efficiency, and prioritize the execution of R&D projects has become an essential topic in theory and practice.

As the project progresses and the workforce structure study becomes more in-depth, the R&D program encounters several challenges, particularly at two levels. Given the limited size of the company and the high availability and mobility of R&D staff, these people, who are the core driving force of R&D, often have to work on multiple projects simultaneously. Practically, whenever a developer moves from a current project to a new one, it takes a period of adjustment to familiarize themselves with the new task, including understanding the context, timeline, and other vital elements. For example, the R&D department of a networking company often develops multiple software suites in parallel [3]. Due to limited resources, it is often necessary to share developers between projects. Employees are often reassigned to other projects after completing their current tasks. However, due to the vastly different business environments in which different software projects operate, these R&D staff involved in the transition may need help seamlessly integrating into their new tasks. In addition, R&D personnel involved in the transfer need to have a deep understanding of the latest task requirements, background information, and the project's current progress [4]. It is worth noting that, unlike the redeployment of resources such as machines and equipment, the movement of personnel between projects is limited by geographic location, which is often difficult to achieve. In contrast to actual employee mobility, transportation time is not directly related to the geographic location of an employee. Such employee mobility can lead to delays in task execution, negatively affecting project progress and efficiency, which needs to be taken seriously by company management.

II. Background of the Study

In 2019, the National Development and Reform Commission (NDRC) listed talent services as one of the key sectors to promote the development of the talent and human capital services industry, which undoubtedly brings significant advantages and unlimited opportunities for the talent services industry to flourish. The core of digital human resource management lies in analytics and forecasting. This system reduces the impact of uncertainty on the workforce and dramatically improves the accuracy of "training and retention" strategies in human resource management [5].

Under human resource management's current challenges, building a highly integrated digital HR system with the organization has become a core mission in the HR field. The HRM system and projects planned in this paper play a pivotal role. In order to optimize the capability of multiple R&D projects in depth, this paper studies the deployment time of R&D personnel between projects in depth. It constructs a model closer to the actual needs based on personnel characteristics [6]. The model not only considers the impact of personnel heterogeneity on task duration but also incorporates the consideration of transfer time and strives to realize the optimal allocation of human resources [7]. In addition, the research results of this paper provide valuable methods and ideas for solving problems in the design of actual R&D projects, with both theoretical depth and practical value. The steps of the random forest algorithm are shown in Table I.

TABLE I. STEPS OF THE RANDOM FOREST ALGORITHM

Random Forest Algorithm
Input: The training dataset has P attributes or predictive variables
Output: Random Forest Algorithm Model for Corresponding Datasets
(1) Check if the decision attribute values in the training dataset are the same and exit if they are the same.
(2) Select feature subset: Randomly select M attributes as prediction variables in the training dataset, where $M \leq P$;
(3) Select the predictive variable with the best classification performance as the root node and decompose it into decision sub-nodes and leaf nodes; and
(4) Repeat steps (2) and (3) to generate N base classifiers.
(5) Average the prediction results for each training tree.

The importance of resource constraints in the planning process has been explored in greater detail in current academic sources on project planning. Relying on resource constraints as the infrastructure of project planning, many resource-constrained challenges can be extended based on real-world contexts [8]. As a critical branch of PSGRC (which may refer to a specific type of project planning or resource constraint problem) research, the topic of multi-project research design has always attracted researchers' attention [9]. Compared to traditional construction projects, R&D projects like software development have a high failure rate. This study aims to ensure the proper allocation of budget and resources to achieve efficient, on-time delivery of products [10]. However, current R&D project planning and human resource optimization strategies must be addressed. Human resources, as a unique renewable resource, are an essential core element in the product development process [11]. In project planning, the time allocation of intangible human resources among project tasks must be fully considered, as well as the possible impact of personnel differences on project planning and personnel deployment strategies.

III. RESEARCH METHODOLOGY

A. *Data Processing*

The data sources were first introduced when constructing the employee turnover prediction model, followed by an in-depth analysis and understanding of the dataset's characteristics. Based on business insights, two key features were successfully created: the number of days absent per year and the actual hours worked per day, respectively [12]. After that, the newly acquired status, survey, and employee performance data were seamlessly integrated into the existing dataset. The integrated data was rationalized into two datasets evaluate the model's performance. Given some of the limitations of the data, a simple and efficient method of integrating measurement units of different natures was used shorten the model's learning and prediction cycle [13]. Data preprocessing steps were also performed, including treating missing values, data substitution, data normalization, and feature selection, which are essential for data integration. The GOSS related process description is shown in Table II.

TABLE II. DESCRIPTION OF GOSS-RELATED PROCESSES

GOSS algorithm description
Inputs: training data, number of iterations d, sampling rate a for extensive gradient data, sampling rate b for small gradient data, loss function, and weak learner
Output: A well-trained, strong learner
(1) it points that have been sorted in descending order, and then randomly select b x (1-a) x 100% sample points from the remaining sample set as a set of minor gradient sample points. Been sorted in descending order, and then randomly select b x (1-a) x 100% sample points from the remaining sample set as a set of minor gradient sample points; and then randomly select b x (1-a) x 100% sample points from the remaining sample set as a set of minor gradient sample points. Points.
(2) Merge the large and small gradient sample sets as the total sample sets for this GOSS sampling.
(3) Multiply the small gradient sample by a weight coefficient (1-a)/b.
(4) Repeat steps (2) and (3) to generate N base classifiers.
(5) Use the sampled samples mentioned above to iteratively generate a new weak learner, repeating it until it reaches the maximum number of iterations or convergence.

Since the raw data was until it reached different files, the processing was broken down into two key steps. First, the incoming data needed to be analyzed in detail, followed by data processing and adding new attributes to this data. Immediately following this, the second step effectively integrates the newly generated data, the survey data, and the employee performance requirements [14]. In the context of this paper, the training process for 4,410 employees is refined into three consecutive phases: analysis of status data, processing of status data, and creation of new status data attributes. Regarding the preprocessing of data, the specific processes are as follows:

1) Data analysis: After in-depth analysis of the data of 4,410 employees over 261 days, it is found that all the employees have been absent for an average of 12 days. The number of days of absence for some of them is even as high as 24 days.

2) Data processing: When all the employees are not present on a particular day, it is regarded as a public holiday, and the relevant test data is directly removed; whereas, when some of the employees have attendance records while the other part does not, the missing value 0 will be filled in the corresponding position.

3) New function creation: The number of absences and the average actual working hours of each employee in 2015 were further calculated. These newly generated attendance, survey, and employee assignment data can now be found in three separate tables.

Given that the preprocessing steps are the same for the training and test packages, the following will focus on the preprocessing process for the training package data. This preprocessing process covers the handling of missing values, data exchange and normalization, and function selection. In the process of random forest or data study, specific attributes often miss values [15]. There are three ways to deal with these missing values: deleting, filling, or leaving them out. If, in practice, the number of functions and data samples are insufficient, it is not recommended to delete these samples directly before providing examples for missing values. Usually, it is necessary to examine the correlation between the missing values and the attribute features to be retained or deleted [16]. There are three broad types of filling of missing values: missing replacement values, missing match values, and model variables. Missing replacement values refer to replacing missing values by filling in statistical indicators or empirical values that do not contain missing data; missing matches correspond to missing values to other characteristics modeled; and model variables refer to data that new, exported, new, or incomplete data have replaced.

Data classification methods in Eq. (1):

$$HC = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}[\left(a_i a_j\right)/(1+l_{ij})]}{A_L^2} \tag{1}$$

It is the sum of the original data and the forward increment of the enterprise.

$$PC = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j P_{ij}^{*}}{A_L^2} \tag{2}$$

In Eq. (2), P_{ij}^{*} the j^{th} probability value of the probability I function is always greater than zero and less than one.

$$RI = 0.5^{*}\, IIC + 0.5^{*}\, PC \tag{3}$$

In Eq. (3), *RI* is the production value of the firm's data, and 0.5 is the coefficient of the two terms.

B. Feature Selection for Random Forests

In data analysis, the use of different units of measurement often leads to very different interpretations of results. In general, attribute values presented in smaller units of measurement tend to appear more prominent, and these values often carry significant importance or "impact." Further, choosing a simplified and functionally meaningful model can help researchers gain a deeper understanding of data generation mechanisms. The feature selection methods can be categorized into three types: filtering, packing, and embedding [17]. This paper uses the correlation coefficient method of the filtering method. This method filters attributes based on the correlation between attributes or setting different thresholds. If the correlation result between two attributes reaches or exceeds the set threshold, it is recognized that there is a strong correlation between them. This paper chose Pearson's correlation coefficient as a correlation measure.

The determination of estimates plays a pivotal role in the selection of critical steps in the experimental process, and this choice is decisive for model selection and the improvement of forecasting accuracy. Since the data in this project are unbalanced and have yet to be analyzed in depth in the balance sheet, the selection of estimates should be closely dependent on the actual data. After careful reading of relevant literature and data, this paper decides to adopt F1, balance accuracy (BA), geometric mean (G-mean), and AUC as the primary evaluation indexes. F1 is a comprehensive score, and the closer its value is to 1, the better the predictive performance of the model is. On the other hand, Balanced accuracy effectively improves the dataset's accuracy by fusing the two metrics, TPR (actual rate) and TNR (actual negative rate). The G-mean, or geometric mean, is the geometric mean of each accuracy level, which effectively filters out the differences in the number of samples in different categories, making the assessment of learning performance more fair. When the ROC curve is closer to the upper left corner, i.e., the area is larger, the AUC value increases, which signals a higher prediction accuracy of the classifier.

Random forest algorithm iteration results:

```
    class A { static void f(){ System.out.println("---------------
-------");
    //} static double quzheng(double a){ int b; int b.
            System.out.println((b=(int)(a+0.5)));
    //return(b);
    } static double qiushang(double a,double b){
    //System.out.println((a/b));
    return(a/b).
    } static boolean odd(int c){
    //if(c%2==0){
    return(false);
    } else{ return(true);
    } } static int juedui(int d){
    //f(d<0){
```

First, the employee turnover prediction model was constructed using a Random forest algorithm by adjusting the parameters before and after optimization. Then, the LXR random forest stack algorithm was further utilized to create a forecasting model for employee turnover. In order to assess the performance of the model entirely, the predictive characteristics of the random forest model before and after optimization were compared and analyzed with the LXR random forest stack model. Simplified methods were introduced to reduce the risk of overload that may arise during the modeling process [18]. These methods include group learning algorithms, traditional data augmentation techniques, and cross-validation methods, which further reduce the risk of model overfitting. In addition, compliance elements were added to construct three specific predictive models: the XGBoost employee intention to leave prediction model, the LightGBM employee intention to leave prediction model, and a two-layer learning logistic regression model based on the LXR stack. It is worth mentioning that the combined LXR stack model was carefully constructed using cross-validation methods to ensure its stability and accuracy. Stacking algorithm model construction process. As shown in Table III.

TABLE III. STACKING ALGORITHM MODEL CONSTRUCTION PROCESS

Stacking algorithm model construction process
Input: training set 1S, base learner H, meta learner H '
Output: Algorithm model of meta learners on the training set
(1) Divide the training set 1S data into K-fold partitions.
(2) Perform K-fold cross-validation using each base learner iH, with K sets of trained data; the
(3) Combine K pieces of data predicted on the training set to obtain new training samples.
(4) Take the average of the predicted data obtained from the test set as the new predicted data.
(5) Import the dataset from step four into the meta learner H classifier (such as logistic regression) to obtain the final prediction result, which is the algorithm model of the meta learner on the training set.

IV. RESULTS AND DISCUSSION

A. Human Resource Effectiveness in Random Forests

Random forest model, as a powerful tool in machine learning, is not only an innovative expansion of traditional data analysis methods but also demonstrates its excellent performance in various data types and complex scenarios. Compared with the traditional decision tree model, Random Forest significantly improves the stability. All this comes from its core algorithm, which aligns with the classic random forest model. In this section, this study specifically focuses on utilizing the Random Forest algorithm to construct a specific HR capability enhancement model, aiming to provide forward-looking guidance for HR planning in an organization by predicting the trend of employee turnover. For this purpose, the Random Forest Classifier tool from the open-source machine learning library Sklearn is the basis for constructing the model.

The setting and tuning of hyperparameters are crucial to constructing the model. For the random forest model, a default starting value for the hyperparameters, i.e., 0, was used as a starting point. This has the advantage that targeted hyperparameter evaluation and tuning can be performed based on the initial performance of the model. However, relying solely on the default starting value often fails to achieve the desired model performance. Therefore, a more refined parameter setting and tuning strategy is used to optimize the Random Forest algorithm model. Among them, FLAML (Fast Lightweight Automated Machine Learning Library) becomes a powerful assistant. FLAML can efficiently tune the hyperparameters in the Random Forest model and find the parameter combinations that can improve the model performance through automation. The questionnaire attributes are shown in Table IV.

Further manual tuning of the hyperparameters was performed on top of the initial optimization of FLAML. This is because, while automated tools can provide a better starting point, more detailed tuning is often required for specific problems and datasets. During the manual tuning process, the model's AUC value (Area Under the Curve) was always used as an evaluation metric to find the combination of hyperparameters that would optimize the model's performance. Overall, a high-performance employee turnover prediction model was successfully constructed by combining Sklearn's Random Forest Classifier tool and the FLAML automated machine learning library. This model can not only help enterprises better understand the pattern of employee turnover but also provide powerful data support for human resource planning to occupy a more favorable position in the fierce competition in the market. LightGBM, the open-source Gradient Boosting Decision Tree (GBDT) algorithmic framework introduced by Microsoft, has excellent parallel learning capability. GBDT, a classic algorithm in the field of random forests, has the core idea of continuously training and correcting the wrong classifiers to achieve the optimal learning effect while avoiding overfitting problems. In addition, GBDT has become a powerful weapon in data research competitions such as Kaggle. In the hyperparameter tuning session, the GBM optical model of FLAML (an automatic machine learning library) is often used to tune the hyperparameters accurately. After completing the initial optimization settings through FLAML, researchers often manually adjust the parameter settings to improve the model's performance further.

TABLE IV. QUESTIONNAIRE ATTRIBUTES

Feature attribute categories	Feature attribute name
Annual attendance data	Employee ID, daily clock-in and clock-out time
Questionnaire survey data	Employee ID, Work Engagement, Last Year's Performance Level, Environmental Satisfaction. Job satisfaction, work-life balance
Work-related detail data	Age, frequency of employee business trips in the previous year, department name, distance from home, education level, education field, number of employees, employee ID, gender, occupational level, professional role, marital status, monthly income, the total number of companies worked for, age 18 or above, salary increase percentage, standard working hours per day, stock options level, total years of work, number of training sessions in the previous year, years of work in the company The following are some examples of the types of training sessions that the company has offered: year, years of work in the company, years since last promotion, and years of working with current direct leaders.

The combined model integrates multiple algorithmic models that have been trained to improve performance while considering the strengths and weaknesses of each model. Combining the ability of primary education students with accurate predictive power strengthens the predictive performance of the overall model. The core modeling strategy can be either a unified model algorithm or a diverse collection of algorithms. Different algorithmic models are usually picked according to the modeling requirements when selecting core strategies. In order to optimize the performance of the stack integration model, the hyperparameter tuning algorithmic model is adopted as the primary base learning tool. During the training and installation process, new learning materials are generated on the first floor of the database learning facility. Directly employing actual data to create new learning records and materials may result in excessive liability risk. Some elements are placed outside the learning area before creating new content to minimize the risk associated with this inconsistency. After parameter optimization, this paper integrates the training methods of these three main machine learning models [19]. using logistic regression to construct an LXR-integrated model to predict employees' intention to leave. This model is used for human resource management in organizations.

In this study, human resources data were divided into two categories, mainly performance and effectiveness evaluation. In the comparison of 18 sample results, pairwise comparisons of the ratios of the two categories were conducted. Negative values were used for the effectiveness evaluation of human resources data to facilitate the comparison of the two data values. The specific results are shown in Fig. 1.

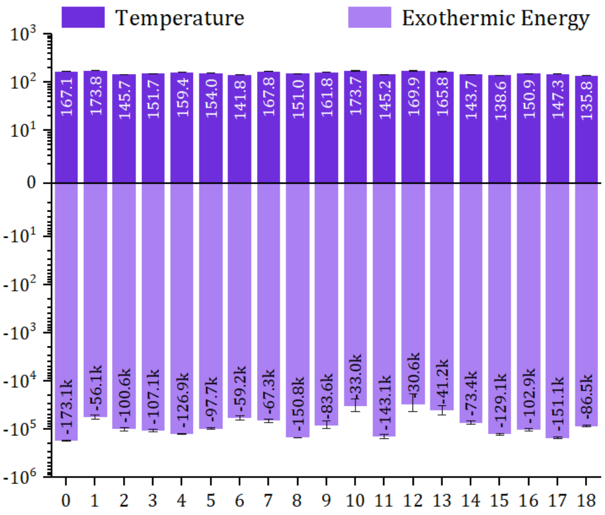

Fig. 1. Human resources data classification.

After going through data integration, missing value filling, data exchange, and standardization, data preprocessing was performed by applying relevant metrics to filter out activities. Three algorithmic models - Random Forest, XGBoost, and Light GBM - were utilized to train the learning toolkit. Given that the second input level is the first core level based on students, the output of the first core level shows a linear correlation with the final classification result. Therefore, a more explanatory model of the logistic regression algorithm was chosen for the second layer of the LXR stack fusion model. After creating new data, the first student used a five-fold cross-validation approach to rank the probability matrix. This strategy effectively reduced the risk of conflict with the second student. When training for Foundation 1 students, optimizing the hyperparameters and subsequently superimposing them is an integral step to obtain superior learning performance. Statistical table for missing feature values in the training set. As shown in Table V.

TABLE V. STATISTICAL TABLE FOR MISSING FEATURE VALUES IN THE TRAINING SET

Attribute Name	Missing quantity	Total number of samples	Missing proportion
Number of companies working	225	8475	0.62%
Total years of service	140	8475	0.65%
Environmental satisfaction	12	8475	0.021%*
Job satisfaction	36	8475	0.63%
Work-Life Balance	552	8475	0.14%

B. *Assessment of Human Resources System Design*

This paper builds a model to predict employees' propensity to leave their jobs. However, the key to making this model truly useful and user-friendly is seamlessly integrating this turnover prediction model into real-world applications. This ensures that employees across the organization can use the model and supports resource decisions. In terms of data protection, this paper adopts the SM2 algorithm to encrypt critical data to safeguard employees' personal information. In addition, the architecture design of the HRMS mentioned in this paper is based on the B/S model while combining the SpringBoot framework and LayUI to ensure the stability and efficiency of the system.

Data construction for enterprise human resources:

```
# Define a function to generate a portion of the ASCII art
def heart_segment(x, y, char).
return ((x - 2 * y) ** 2 - (x ** 2 - 2 * x * y + y ** 2) ** 2 / 25).substitute(x=x, y=y).replace('**', '^') <= char
def print_heart(char):
for y in range(10):
for x in range(20):
import turtle
import math
turtle.pen()
t=turtle
t.up()
t.goto(0,150)
t.down()
t.color('red')
t.begin_fill()
if heart_segment(x / 4.0, y, char).
print(char, end='')
else.
print(' ', end=' ')
print()
print_heart('*')
```

Before embarking on system development, in-depth research and careful consistency analysis occupy a pivotal position in the system construction process. In this paper, through detailed market research, we have explored users' actual needs and clarified the system architecture's core value and far-reaching significance. In this process, functional and non-functional requirements are explored in detail, and the specific content of system requirements is finally established. Based on related research, it was identified that system users are mainly categorized into three roles: ordinary employee users, employee administrator users, and system administrator users. The tasks commonly involved in the daily operations of these three roles include logging into the system, logging out of accounts, and changing passwords. Regular employee users may also have access to pay for paychecks and online evaluations. The rating management segment focuses on maintaining and managing rating names, statuses, and flags. Employees can participate only when the assessment mode is activated. The department management module provides users with the ability to store information in their department; the user management module allows authorized users to enter and update information about system users; the role management module not only enables users to manage roles within the system but also ensures that the permissions of each user role are appropriately maintained; and finally, the menu management module is responsible for assigning authorized users the names, paths, and other essential information of the system's menus. Finally, the menu management module specifies the system menus' names, paths, and other essential information for authorized users. The parameters of the random forest are described in Table VI.

TABLE VI. DESCRIPTION OF RANDOM FOREST PARAMETERS

Parameter	Parameter meanings	Default value
(an official) standard	The function that measures the quality of segmentation, with options such as gini, entropy, and log_loss	5565
greatest feature	The size of the random subset of features is considered when dividing nodes. The lower the value, the more the variance decreases, but the more the deviation increases	1124
largest node	The maximum value of leaf nodes, used to limit their growth	12
incremental	The number of base learners is usually better with larger values, but the computation time increases accordingly. When the number of trees exceeds a When the number of trees exceeds a critical value, the performance of the algorithm does not significantly improve.	1
random number	Used to control the randomness of samples during tree construction	6

In this paper, the HR architecture consists of three main layers: the data layer, the control layer, and the interaction layer (i.e., the user interface layer). The data layer is responsible for storing, maintaining, and protecting the system data. The control layer is responsible for receiving requests from the interaction layer, interacting with the data layer, and finally sending the processed results back to the interaction layer. The interaction layer is committed to providing users with an intuitive and easy-to-understand web interface, which transforms complex business logic into a visual interface. This interface allows users to quickly realize all kinds of business needs and form an intuitive user experience. With the rapid development of artificial intelligence technology and the increasingly refined division of labor in society, enterprises have put forward more stringent requirements and cost control standards for the digital management of human resources. In order to quickly adapt to the changing operating environment and market demand, companies must continuously optimize their management and internal processes. Therefore, using the Random Forest method to predict employee turnover has become an effective strategy for HR managers to reduce personnel risks and costs. The random forest algorithm with random correction is shown in Table VII.

TABLE VII. RANDOM FOREST ALGORITHM WITH STOCHASTIC CORRECTION

Variant	Search (1)	Search (2)	Envir (1)	Envir (2)
modified effect	0.1417***	0.0214***	0.0654***	0.0251***
observed value	5147	5147	5147	5147
control variable	be	be	be	be
Annual fixed effects	be	be	be	be
industry fixed effect	be	be	be	be
area fixed effect	clogged	clogged	clogged	be

A specific model for enterprise data effectiveness assessment:

$$X_{new} = x_i + rand(0,1) \times (y_j - x_i)(j = 1,..., P) \quad (4)$$

In Eq. (4) $rand(0,1)$ is the reordering of the independent variables of function x; $Mar(Q,V_T)$ the function is specified as follows:

$$Mar(Q,V_T) = ave(V_T) - E^* \le \frac{\overline{\rho}(1-s^2)}{\delta^2} \quad (5)$$

$$F1 = \frac{2 \times precision \times recall}{precision + recall} \quad (6)$$

In Eq. (5) and Eq. (6), $ave(V_T)$ is the functional form for finding the mean of the *V* function. It $precision \times recall$ is the precision and return value of the function *F1*.

The process of product creation is highly personalized and intelligent. Unlike traditional industrial products, which rely on repetitive mechanical operations, it relies heavily on the ingenuity and innovation of researchers and developers. During the project preparation stage, the R&D staff enjoys the support of abundant resources, such as various types of machinery and equipment. The diversity of human resources refers to the differences in technical expertise, knowledge base, and personal attributes of the company's employees. This diversity has two significant effects on staff allocation: first, the diversity of staff skills directly affects the timeliness of task completion, which is reflected in the different skills of researchers and developers with the same qualifications in terms of age, seniority, experience, and education, which in turn has an impact on performance. Second, employee diversity is also affected by the time between multiple projects. In the actual R&D process, each developer is unique in terms of the underlying design experience he or she possesses. Taking the R&D department of a Web enterprise as an example, any software project can be disassembled into several independent tasks, each carrying a different workload and content, during the functional analysis and requirements combing phase before the official launch. From the perspective of task allocation, equipping a single task with a combined team with different R&D skills and experience can effectively adjust the project delivery circle.

In the data comparison of Fig. 1, it was found that the performance evaluation data had more significant differences than the performance evaluation data. Therefore, performance evaluation data should be selected for analysis of data density in the study. Fig. 2 shows the process of data density analysis. In order to more significantly highlight the results of the data difference test, polar coordinates were used for this test. The data was divided into intervals of 0-2 π and subjected to extreme segmentation approaching "positive infinity", resulting in a continuous polar coordinate graph of the stability level of the data. The results showed that only at 1/8-3/8 of the data, the stability exceeded the LV3 level, but at other stages, the data exceeded the LV1 level, meeting the testing criteria, as shown in Fig. 2.

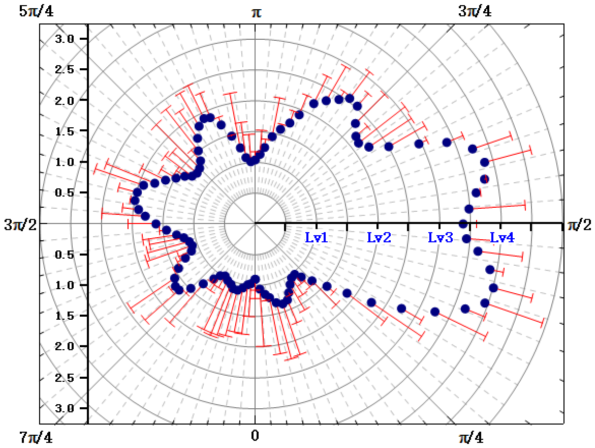

Fig. 2. Human resources effectiveness assessment density.

V. Conclusion

This study has achieved a series of significant findings and results through the classification and effectiveness assessment of enterprise human resource data based on the random forest algorithm. With the continuous development of information technology, effective management of human resources in enterprises has become increasingly important, and the method proposed in this study provides an effective way to process and analyze data, providing deeper insights and decision support for human resource management. First of all, by classifying enterprise human resource data, it is possible to manage and analyze employees more refinery. The classification model constructed based on the Random Forest algorithm can accurately categorize employees according to their characteristics and labels, helping enterprises understand the basic situation of employees, their work characteristics, and potential development direction. This gives enterprises a more comprehensive view of talent management, which helps them develop more personalized training, motivation, and promotion plans for different groups and improves employee job satisfaction and loyalty. Secondly, the performance evaluation model provides an objective and scientific performance evaluation method for enterprises. Analyzing employee performance data and other relevant indicators can assess the contribution and effectiveness level of employees in the organization and help companies identify high-performance employees and potential room for performance improvement. This provides an essential decision-making basis for enterprises and helps optimize the allocation of human resources and improve the overall performance and competitiveness of the organization.

The classification and effectiveness assessment method of enterprise human resource data based on the random forest algorithm has important theoretical significance and practical value. The automated processing and analysis of enterprise human resource data can better understand the characteristics and behaviors of employees and provide more accurate and comprehensive decision support for enterprise managers. In the future, the algorithmic model can be further optimized, combined with more data sources and indicators, and the classification and evaluation system can be improved to cope with the ever-changing market environment and enterprise needs. At the same time, the model can be validated and applied in combination with specific business scenarios and actual cases to improve its accuracy and practicality further and provide more robust support for the sustainable development and innovation of enterprises.

Acknowledgment

This work is supported by Funding project for the construction of disciplines and specialties of Henan Private Ordinary Higher Education Institutions in 2023， " business administration" (Jiaozhengfa [2022] No. 377).

References

[1] Teles, G., Rodrigues, J. J., Rabelo, R. A., & Kozlov, S. A. (2021). Comparative study of support vector machines and random forests machine learning algorithms on credit operation. *Software: Practice and Experience*, 51(12), 2492–2500. https://doi.org/10.1002/spe.2842

[2] Assiri, A. (2021). Anomaly classification using genetic algorithm-based random forest model for network attack detection. *Computers, Materials & Continua*, *66*(1). https://doi.org/10.32604/cmc.2020.013813

[3] Nasar, N., Ray, S., Umer, S., & Mohan Pandey, H. (2021). Design and data analytics of an organization's electronic human resource management activities through the Internet of Things. *Software: Practice and Experience*, *51*(12), 2411–2427. https://doi.org/10.1002/spe.2817

[4] Bazzaz Abkenar, S., Mahdipour, E., Jameii, S. M., & Haghi Kashani, M. (2021). A hybrid classification method for Twitter spam detection based on differential evolution and random forest. *Concurrency and Computation: Practice and Experience*, *33*(21), e6381. https://doi.org/10.1002/cpe.6381

[5] Tassi, A., Gigante, D., Modica, G., Di Martino, L., & Vizzari, M. (2021). Pixel-vs. Object-based Landsat 8 data classification in Google Earth engine using random forest: The Maiella National Park case study. *Remote Sensing*, *13*(12), 2299. https://doi.org/10.3390/rs13122299

[6] Pallathadka, H., Ramirez-Asis, E. H., Loli-Poma, T. P., Kaliyaperumal, K., Ventayen, R. J. M., & Naved, M. (2023). Applications of artificial intelligence in business management, e-commerce, and finance. *Materials Today: Proceedings*, pp. *80*, 2610–2613. https://doi.org/10.1016/j.matpr.2021.06.419

[7] Chaudhary, M., Gaur, L., Jhanjhi, N., Masud, M., & Aljahdali, S. (2022). Envisaging employee churn using MCDM and machine learning. *Intelligent Automation & Soft Computing*,*33*(2). https://doi.org/10.32604/iasc.2022.023417

[8] RL, M., & Mishra, A. K. (2022). Measuring financial performance of Indian manufacturing firms: Application of decision tree algorithms. *Measuring Business Excellence*, *26*(3), 288–307. https://doi.org/10.1108/MBE-05-2020-0073

[9] Javed Mehedi Shamrat, F., Ranjan, R., Hasib, K. M., Yadav, A., & Siddique, A. H. (2022). Performance evaluation among id3, c4. 5, and cart decision tree algorithm. *Pervasive Computing and Social Networking: Proceedings of ICPCSN 2021*, pp. 127–142. https://doi.org/10.1007/978-981-16-5640-8_11

[10] Siahaan, M. (2021). An analysis of contract employee performance assessment using machine learning. *Journal Of Informatics And Telecommunication Engineering*, *5*(1), 121–131. https://doi.org/10.31289/jite.v5i1.5357

[11] Chen, Y., Zheng, W., Li, W., & Huang, Y. (2021). Large group activity security risk assessment and risk early warning based on random forest algorithm. *Pattern Recognition Letters*, pp. *144*, 1–5. https://doi.org/10.1016/j.patrec.2021.01.008

[12] Zhang, T., Su, J., Xu, Z., Luo, Y., & Li, J. (2021). Sentinel-2 satellite imagery for urban land cover classification by optimized random forest classifier. *Applied Sciences*, *11*(2), 543. https://doi.org/10.3390/app11020543

[13] Chen, Y., Chen, W., Chandra Pal, S., Saha, A., Chowdhuri, I., Adeli, B., Janizadeh, S., Dineva, A. A., Wang, X., & Mousavi, A. (2022). Evaluation efficiency of hybrid deep learning algorithms with neural network decision tree and boosting methods for predicting groundwater potential. *Geocarto International*, *37*(19), 5564–5584. https://doi.org/10.1080/10106049.2021.1920635

[14] Yahia, N. B., Hlel, J., & Colomo-Palacios, R. (2021). From big data to deep data to support people analytics for employee attrition prediction. *IEEE Access : Practical Innovations, Open Solutions*, *9*, 60447–60458.

[15] Es-Sabery, F., Es-Sabery, K., Qadir, J., Sainz-De-Abajo, B., Hair, A., García-Zapirain, B., & De La Torre-Díez, I. (2021). A MapReduce opinion mining for COVID-19-related tweets classification using enhanced ID3 decision tree classifier. *IEEE Access : Practical Innovations, Open Solutions*, *9*, 58706–58739.

[16] Zhu, H. (2021). Research on human resource recommendation algorithm based on machine learning. *Scientific Programming*, *2021*, pp. 1–10.

[17] Farhadi, H., & Najafzadeh, M. (2021). Flood risk mapping by remote sensing data and random forest technique. *Water*, *13*(21), 3115. https://doi.org/10.3390/w13213115

[18] Garg, S., Sinha, S., Kar, A. K., & Mani, M. (2022). A review of machine learning applications in human resource management. *International Journal of Productivity and Performance Management*, *71*(5), 1590–1610. https://doi.org/10.1108/IJPPM-08-2020-0427

[19] Zhang, Y., Xu, S., Zhang, L., & Yang, M. (2021). Big data and human resource management research: An integrative review and new directions for future research. *Journal of Business Research*, *pp. 133*, 34–50. https://doi.org/10.1016/j.jbusres.2021.04.019