Theory and Practice of Science and Technology
2025, VOL. 6, NO. 6, 92-101
DOI: 10.47297/taposatWSP2633-456919.20250606

Research on the Prediction of First-Line Workers' Turnover Risk in Auto Parts Manufacturing Companies Based on Catboost

Xiao Xiao[1], XueYang Zhong[2]*

[1]School of Humanities and Education, Guangzhou Institute of Science and Technology,Guangzhou, 510540, China

[2]School of Computer Science and Engineering, Guangzhou Institute of Science and Technology,Guangzhou, 510540, China

ABSTRACT

In the highly competitive global automotive industry, the stability of front-line workers in auto parts manufacturing enterprises is vital for the industry's supply chain, but their high turnover rate has become a bottleneck restricting enterprises' sustainable development. This study constructs a front-line worker turnover risk prediction model for a South China auto parts manufacturer based on the CatBoost algorithm. By collecting 28 features from multiple management systems and expert evaluations, performing data preprocessing, hyperparameter tuning, and model comparison, the CatBoost model shows significant advantages in handling categorical features and preventing overfitting. Its prediction accuracy reaches an AUC of 0.92 and an F1 score of 0.886 after oversampling, far surpassing traditional models such as logistic regression and random forest. Feature importance analysis reveals that factors such as job position and seniority significantly impact turnover risk. Targeted interventions based on model predictions have reduced the quarterly turnover rate from 6.3% to 4.5%, and in peak periods from 8.3% to 6.1%. This research provides an innovative and practical solution to address the turnover of front-line workers, with future studies to further optimize the model and verify its long-term effectiveness.

KEYWORDS

Auto parts manufacturing enterprises; CatBoost algorithm; First-line workers; Turnover risk prediction

1 Introduction

In the current highly competitive global automotive industry, the auto parts manufacturing industry serves as the foundation of the automotive industry chain, its stability being critical to ensuring supply chain resilience. First-line workers, as the core workforce in production operations, directly influence enterprise efficiency, product quality, and cost control [1-2]. However, persistent high turnover rates among front-line workers have become a significant challenge for auto parts manufacturers [3]. Specifically, data from a South China-based auto parts company (2022-2023), the research subject of this article, reveal a turnover rate exceeding 20%, which has affected production continuity and market competitiveness. This issue not only inflates labor costs, but also leads to productivity declines and quality inconsistencies, highlighting the urgent need for accurate prediction of turnover risk [4-5].

Existing research on employee turnover has taken advantage of theoretical frameworks from human resource management (e. g. career development theory, family life cycle theory) and diverse methodologies, ranging from traditional logistic regression to machine learning techniques such as random forests [6-9]. However, these approaches face limitations in auto parts manufacturing scenarios, including poor handling of high-dimensional categorical data, overfitting risks, and inadequate consideration of industry-specific data collection constraints [10-11].

This study addresses these gaps by focusing on a leading South China auto parts manufacturer. By integrating 28 characteristics from both internal management systems (for example, HR, production, equipment) and expert assessments, we develop a CatBoost-based prediction model tailored to the industry ' s operational characteristics. The CatBoost algorithm is selected for its superior capabilities in categorical variable processing, over-fit mitigation, and optimization of high-dimensional data modeling [12]. Through this research, our objective was to develop a robust turnover risk prediction model for front-line workers in auto parts manufacturing, validate the superiority of CatBoost versus traditional models (e. g., logistic regression, random forest) in prediction accuracy, and identify key influencing factors to inform targeted HR interventions to improve workforce stability and operational efficiency. By bridging theoretical insights with industry-specific data practices, this study aims to provide a scalable solution to proactive turnover management in the auto parts sector, contributing to both academic literature and practical HR strategies.

2 Principles and advantages of the catboost algorithm

2.1 Algorithm principles

The CatBoost algorithm is a machine learning algorithm developed based on the Gradient Boosting Decision Tree (GBDT)framework. Its core mechanism is to gradually fit the gradient of the objective function through an iterative method to construct a prediction model. In the iterative process of increasing the gradient, each iteration builds a new

decision tree to fit the residual errors of the previous iteration. Specifically, assume that (m- 1) iterations have been completed and the model obtained at this time is $F_{m-1}(x)$. Then, in the m-thiteration, the algorithm calculates the negative gradient of the current model. The calculation formula is:

$$R_{m,i} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)} \tag{1}$$

Here, $L(y_i, F(x_i))$ is the loss function, which is mainly used to measure the degree of difference between the predicted value $F(x_i)$ and the true value y_i. The goal of the newly constructed decision tree $h_m(x)$ is to fit these negative gradients, making the prediction result of the model in this step closer to the true value. The final model $F_m(x)$ is composed of the weighted sum of all previous decision trees:

$$F_m(x) = F_{m-1}(x) + \lambda h_m(x) \tag{2}$$

Here, λ is the learning rate, balancing model convergence speed and prediction accuracy .A defining feature of CatBoost is its handling of categorical features through **Ordered Target Statistics (OTS)** encoding. Unlike traditional one-hot encoding, OTS leverages target variable statistics to en-code categorical values without manual preprocessing. For a categorical feature value c, the encoding is computed using historical data from samples preceding the current instance in a random permutation:

$$TS(c) = \frac{\sum_{j \in history(i), x_j^c = c} y_j + a \cdot p}{\sum_{j \in history(i), x_j^c = c} 1 + a} \tag{3}$$

Here, history(i) represents the set of samples that come before i in the permutation, a is a smoothing parameter, and p is the global prior probability of the target [13]. This method avoids target leakage by excluding the current sample ' s label, enabling the model to capture non-linear relationships between categorical features (e.g., job position, department) and turnover risk more effectively.

CatBoost further incorporates symmetric gradient boosting, which randomizes data permutations across iterations to stabilize gradient estimation and reduce overfitting. By averaging gradients computed under different orderings, the algorithm mitigates bias from the data sequence, making it particularly robust for high-dimensional datasets with imbalanced labels, as seen in employee turnover data from automotive parts manufacturing [14]–[16].

2.2 Algorithm advantages

2.2.1 Efficient handling of categorical features

Unlike other machine learning algorithms, the CatBoost algorithm does not require complex preprocessing of categorical features, such as one-hot encoding. Its built-in encoding methodcandirectlyprocesscategoricalcharacteristics, which not only saves a large amount of computing resources and time, but also better captures the internal relationship between categorical characteristics and the target variable [15]. In this research, categorical features, such as the department where employees are located and the type of job, can be directly input into the training model. This method greatly simplifies the data preprocessing process, improves the learning ability of the model, and thus improves the accuracy in predicting the turnover risk of first-line workers [16].

2.2.2 Strong anti-overfitting ability

TheCatBoost algorithm adopts a variety of techniques to prevent overfitting. Among them, the symmetric gradient boosting technology makes the gradient estimation more stable, avoiding the model ' s over-reliance on specific data patterns during the training process. At the same time, by randomly permuting and combining the training data, the diversity of the data is increased, allowing the model to learn a wider range of data features instead of being limited to certain local features. This feature enables the model to effectively avoid overlearning the noise in the data when dealing with complex problems such as predicting the risk of turnover of first-line workers in auto parts manufacturing enterprises, resulting in more reliable prediction results and improving the generalizability of the model [15][16].

2.2.3 Fast training speed and excellent performance

The CatBoost algorithm adopts advanced techniques such as the histogram algorithm in its algorithm design, which can quickly process large-scale data. During the training process, the CatBoost algorithm discretizes continuous features in a histogram form, which greatly reduces the amount of calculation and speeds up the model training speed. Benefiting from its outstanding advantages in processing categorical features and anti-overfitting, the CatBoost algorithm also outperforms similar algorithms in prediction accuracy. The experimental results show that in the task of predicting the risk of turnover of employees in this research, the precision, recall rate, and F1 score of the CatBoost model are higher

than those of traditional models such as logistic regression and random forest, demonstrating its high efficiency and superiority [16-18].

3 Research design

3.1 Research objectives

This research aims to build a turnover risk prediction model for first-line workers based on the CatBoost algorithm using data that can be obtained from the daily management of auto parts manufacturing companies, and compare and analyze it with commonly used models such as logistic regression and random forest. By comparing the performance of each model in terms of prediction accuracy, recall rate, F1 score, and area under the ROC curve (AUC) [19], the advantages and applicability of the CatBoost model in this field are thoroughly evaluated. Using the built-in CatBoost model, enterprises can accurately identify employees with a high turnover risk in advance, providing a strong basis for enterprises to formulate scientific and effective human resource management strategies.

3.2 Research subjects

This research selects first-line workers from an auto parts manufacturing enterprise in South China, China as research subjects. The research data covers all first-line workers during the two years from April 1, 2022, to March 31, 2023, with a total of 582 employees. It includes both employees who left the job during these two years and newly joined employees. The data of these employees form the basis of this research.

3.3 Data collection

The data comes from multiple management systems in the enterprise, including the Human Resources (HR) management system, the production management system, the attendance system, and the equipment management system. Data on 28 features of 582 employees were obtained from these systems (see Table 1). These data cover multiple dimensions such as employees' personal situations and job data and are all objective data. In addition, through the evaluation of an internal expert group of the company, subjective data such as the number of process steps, the requirements for job skill level, the difficulty in job quality control, the severity of the consequences of incorrect operations, and the requirements for internal team collaboration were obtained. To ensure scientificity and standardization of the evaluation, these subjective data are graded from 1 to 5, and the evaluation criteria for each level are determined by the internal expert group of the company based on professional standards and practical experience. During the two-year research period, these characteristics were updated quarterly and a total of 4,586 pieces of data were generated. In this research, whether an employee leaves the job in the next quarter after data collection is used as the dependent variable, with "1" indicating leaving the job and "0" indicating not leaving the job [19]–[21].

Table 1 Selected Features

Feature	Resource	Feature	Resource
age	HR Management System	process_steps_count	Expert assessment
gender	HR Management System	skill_level_requirement	Expert assessment
education_level	HR Management System	quality_control_difficulty	Expert assessment
marital_status	HR Management System	error_operation_consequence_severity	Expert assessment
department	HR Management System	team_internal_collaboration_requirement	Expert assessment
position	HR Management System	team_member_stability	HR Management System
work_experience	HR Management System	job_transfer_count	HR Management System
job_tenure	HR Management System	equipment_failure_time	Equipment Management System
monthly_salary	HR Management System	equipment_failure_count	Equipment Management System
overtime_hours	HR Management System	last_punishment_duration	HR Management System
performance_score	HR Management System	distance_home_work	HR Management System
attendance_rate	HR Management System	last_promotion_duration	HR Management System
training_frequency	HR Management System	number_of_projects	Project Management System

3.4 Basis for feature selection

When constructing a turnover risk prediction model for first line workers in auto parts manufacturing companies, this research selected 28 characteristics. These features mainly cover categories such as employee personal information, job characteristics, work performance and achievements, and the impact of the enterprise environment. The selection of features considered comprehensively the theoretical basis, the actual needs of the organization and the availability of

data to ensure the scientificity, practicality, and feasibility of the research.

From a theoretical perspective, relevant theories in human resource management and organizational behavior provide a basis for feature selection. For example, career development theory and family life cycle theory emphasize that the characteristics of employees at different stages can affect their turnover decisions. Therefore, features related to personal information from employees, such as age and marital status, are included. These features help to explore the impact on turnover risk from the perspectives of personal life and career development [11].

Combined with the actual operation needs of enterprises, the features in the categories of job characteristics, work performance and achievements, and the impact of the enterprise environment are crucial. The characteristic characteristics of the job, including the type of job, the number of process steps and the requirements of the job skill level, etc., the differences in the content of the work, the skill requirements, and the prospects of development of different jobs will directly affect the willingness of the employees to stay or leave [11]. Work performance and achievement characteristics, such as performance scores and the number of projects participated in, reflect the quality of the work of the employees [19]. Employees with poor work results or who do not meet expectations have a relatively higher turnover risk [22]. The impact features of the enterprise environment, such as the failure time of the equipment, the failure count, the atmosphere of the department, etc., the stability of production equipment and the internal environment of the enterprise, will affect the work experience of the employees and thus affect the turnover risk.

Factors such as organizational identification and employee satisfaction, which theoretically affect turnover risk, were not included in this study because they are difficult to obtain in the daily management of enterprises. The features selected in this research have stable and convenient acquisition channels, greatly reducing the difficulty and cost of data collection, and have a high degree of feasibility and practicality in actual operation and application [22].

3.5 Data preprocessing

In the research to predict the risk of turnover of first-line workers in auto parts manufacturing enterprises using the CatBoost model, the original data often has problems such as incompleteness, high noise, and complex data types. If these original data are used directly for model training, it will seriously affect the performance and prediction accuracy of the model. Therefore, it is necessary to preprocess the original data and convert them into a suitable format to input into the CatBoost model. Specific steps include data cleaning, data standardization, feature selection, and categorical feature encoding [23].

3.5.1 Data cleaning

After a preliminary review of the data, it was found that there are minimal missing data and that they can be supplemented manually. Therefore, no extensive data filling operations are required.

3.5.2 Categorical feature encoding

Although the CatBoost model can directly process categorical characteristics, this research performed appropriate encoding of some categorical characteristics to further improve training efficiency and model performance. For categorical features with an order relationship, such as employees' job ranks and educational levels, the Label Encoding method was used to map each category to an integer, thereby retaining the order information between categories. For categorical features without an order relationship, such as the department where employees are located and job types, since the CatBoost model internally adopts a special Ordered Target Statistics encoding (also known as Ordered TS encoding) method. This encoding method combines the information from the target variable and can effectively process categorical features. There is no need to perform additional encoding processing, such as one-hot encoding on categorical features, in advance, like some other machine learning models. They can be directly inputted into the model for training [17-18].

3.5.3 Data standardization

Although the CatBoost model is insensitive to the scale of features, to improve the convergence speed and stability of the model, this research standardized the numerical features. The Z score standardization method was applied, transforming each numerical feature into a standard normal distribution (mean=0, standard deviation=1)[24]. The specific calculation formula is:

$$Z = \frac{X - \mu}{\sigma} \tag{4}$$

3.5.4 Data imbalance problem

The research data is updated quarterly, and a total of 4596 pieces of data were accumulated in two years. Among them, there are 223 pieces of data marked as "left the job" and 4373 pieces of data marked as "did not leave the job". The annual turnover rate of the company's front-line employees in two years is approximately 20%. However, during the data collection process, the data of employees who left the job are no longer collected, while the data of employees who did not leave the job are collected quarterly. This leads to a low proportion of data marked as "left the job" in the dataset, and there is a significant difference in the number of samples.This data imbalance will cause the CatBoost model to be biased toward the majority class during the training process, resulting in poor learning of the minority class (employees who left the job) samples and affecting the prediction accuracy of the model for employees who left the job. In this study, the SMOTE algorithm (Synthetic Minority Oversampling Technique) was used for oversampling to expand the data of employees who left the job [24]. The specific steps are as follows: First, calculate the distance between each sample x in the minority class (samples of employees who left the job) and other samples, and obtain the k nearest neighbors through the KNN (K-Nearest Neighbors) algorithm. Then, based on the data imbalance ratio, set the sampling ratio to determine the sampling multiplier, and randomly select a certain number of samples (denoted x') from the k nearest neighbors of x. Finally, new samples can be generated through the formula [25]:

$$x_{new} = x + rand(0,1) \times (x' - x) \tag{5}$$

After processing the talent data imbalance problem with the SMOTE method in this study, the number of employees who left the job and those who did not leave the job is 4373 person- times each, alleviating the data imbalance problem.

3.6 Selection of comparative models

To comprehensively evaluate the effectiveness of the CatBoost algorithm in predicting the turnover risk of front-line workers in automotive parts manufacturing enterprises, this study selected multiple classic machine learning models for comparison with CatBoost [9-15]. The specific selection basis is as follows:

3.6.1 Traditional statistical model

Logistic regression was selected as the representative of linear classification models. Based on probability theory, this model has good interpretability and is often used to analyze factors that influence employee turnover, being widely applied in the field of human resources. Including it in the comparison helps to verify the advantages of CatBoost in non-linear relationship modeling [9,20].

3.6.2 Ensemble learning models

Random Forest and LightGBM, as typical representatives of ensemble learning, were selected. The Random Forest can effectively reduce the risk of overfitting by building multiple decision trees and integrating the results. LightGBM adopts the histogram algorithm and Leaf- wise growth strategy, showing high efficiency in processing large-scale data. Comparison with CatBoost can highlight the unique advantages of CatBoost in handling categorical features and preventing overfitting [15,20].

3.6.3 Instance-based and probability-based models

KNN(K-Nearest Neighbors algorithm) is an instance- based non- parametric learning method that classifies by calculating the distance between samples. Naive Bayes is based on Bayes' theorem and makes probability predictions assuming feature conditional independence. The principles of these two types of models are quite different from those of CatBoost, which helps to verify the prediction ability of the CatBoost model under complex data features from multiple perspectives [9,20].

By comparing the different types of models mentioned above, this study will evaluate the performance of the CatBoost model from multiple dimensions, providing a more persuasive method reference for predicting the turnover risk of front-line workers in auto-parts manufacturing enterprises.

3.7 Hyperparameter settings of the catboost model

When constructing the turnover risk prediction model for front- line workers in auto parts manufacturing enterprises, the hyperparameter settings of the CatBoost model play a crucial role in its performance. In this study, through fivefold cross-validation, the hyperparameters were systematically tuned based on the F1 score and the AUC index of the validation set [14,16](see Table 2).

Table 2 Hyperparameter settings of the CatBoost model

Hyperparameter	Value	Parameter Meaning and Setting Basis
learning_rate	0.05	Controls the step size of model parameter updates in each iteration. Through testing, a learning rate of 0.05 can ensure the rapid convergence of the model while effectively avoiding overfitting caused by too large a step size.
iterations	500	Determines the total number of integrated decision trees. As the number of trees increases, the model's fitting ability enhances. However, after exceeding 500 trees, the performance of the validation set tends to be stable. Therefore, this value is selected to balance complexity and generalization ability.
depth	6	Limits the maximum number of levels of a single decision tree. When the depth is 6, the model can not only learn complex patterns in the data but also avoid overfitting problems caused by overly deep trees.
Categorical Feature Processing	-	Using Ordered Target Statistics encoding with a smoothing parameter set to 3. This setting effectively reduces the impact of data sparsity and fully explores the potential relationship between categorical features such as departments and job types and turnover risk.
random_seed	42	Fixing the random seed ensures the reproducibility of the experiment, making the model training process consistent and facilitating subsequent verification and comparative analysis.
early_stopping_rounds	20	When the validation set indicators do not show significant improvement for 20 consecutive iterations, the training stops automatically to prevent overfitting and improve training efficiency.

3.8 Hyperparameter settings of other comparative models

To ensure the fairness and rigor of the comparative experiment, this study tuned the hyperparameters of each comparative model [26-27] .(see Table 3)

Table 3 Hyperparameter Settings of Other Comparative Models

Model Name	Key Hyperparameters	Values
Logistic Regression	penalty, C, solver, max_iter	L2, 1, liblinear, 1000
Random Forest	n_estimators, max_depth, min_samples_split, max_features	100, None, 2',sqrt'
Support Vector Machine (SVM)	kernel, C, gamma, probability	RBF, 1',scale', TRUE
XGBoost	learning_rate, max_depth, n_estimators, subsample, objective	0.1, 6, 100, 0.8, 'binary:logistic'
K-Nearest Neighbors (KNN)	n_neighbors, weights, metric	5, 'uniform', 'euclidean'
Naive Bayes	alpha, priors	1, None

4 Employee turnover prediction

After completing preprocessing processes such as data cleaning, standardization, feature encoding, and imbalance handling, this study adopted a 75%- 25% data splitting strategy to divide the dataset into a training set and a test set. Based on this, an employee turnover prediction model with the CatBoost algorithm as the core was constructed and systematically compared with classic classification models such as Logistic regression, Random Forest, KNN, Naive Bayes and LightGBM.

In the research, first, the traditional models such as Logistic regression were tuned and trained. The hyperparameter con-figuration was optimized through five-fold cross-validation, and the performance of each model in terms of precision, recall rate, F1-score, and AUC was evaluated based on the test set. Subsequently, the CatBoost algorithm was used to construct a prediction model. Under the same dataset division and evaluation criteria, the prediction performances of different models were compared and analyzed. At the same time, using the CatBoost model's feature importance evaluation mechanism, the 28 characteristics that affect the risk of turnover of employees were ranked by weight to identify the key influencing factors, providing a quantitative basis for companies to formulate precise intervention strategies[12].

4.1 Comparative analysis on the original dataset

The performance indicators of the 6 models were evaluated on the original dataset, and the specific results are shown in Table 4.

After processing the dataset using the SMOTE algorithm for oversampling, the performance indicators of the 6 models were evaluated again, and the results are shown in Table 5 below.

Through the data comparison of the two tables, it can be seen that the prediction effect of the models has been greatly improved after oversampling with the SMOTE algorithm.

Table 4 Performance indicators on the Original Dataset

Model	Precision	Recall Rate	F1 score
Logistic Regression	0.683	0.726	0.704
Random Forest	0.746	0.669	0.705
KNN	0.751	0.825	0.786
Naive Bayes	0.749	0.763	0.756
LightGBM	0.795	0.782	0.788
CatBoost	0.822	0.802	0.812

Table 5 Performance indicators on the processed l Dataset

Model	Precision	Recall Rate	F1 score
Logistic Regression	0.782	0.768	0.775
Random Forest	0.795	0.738	0.765
KNN	0.817	0.841	0.829
Naive Bayes	0.794	0.846	0.819
LightGBM	0.804	0.812	0.808
CatBoost	0.896	0.877	0.886

4.2 ROC curve and auc value analysis

On the preprocessed dataset, the ROC curves of each model were drawn respectively[14], as shown in Figure. 1.

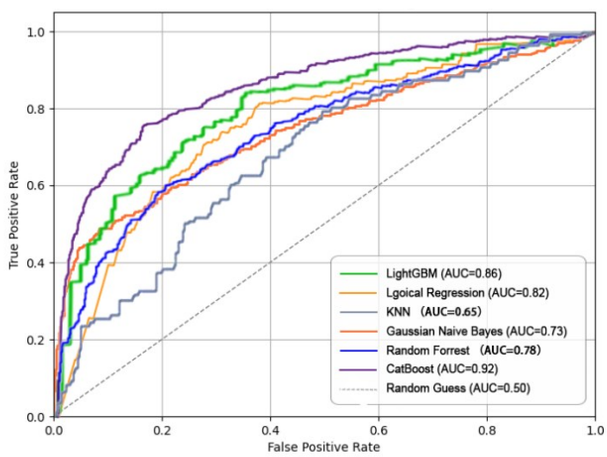

Figure. 1 The ROC curves of each model

This figure intuitively shows the classification effects of the 6 models after oversampling, and the performance of the models was compared through the AUC values. This evaluation process was carried out based on the test - set data, aiming to verify the generalization ability of the models, that is, the performance of the models on unseen data.

The KNN model and the Naive Bayes model showed relatively weak performance in this employee turnover prediction task. As an instance - based learning algorithm, the KNN model mainly relies on the distance between samples for classification decisions when dealing with complex problems such as employee turnover prediction. However, employee turnover is affected by multiple complex factors, and the data distribution is not a simple geometric structure. This makes it difficult for the KNN model to accurately capture the internal laws. Therefore, its AUC value is only 0.65, indicating that the KNN model has great limitations in distinguishing between employees who leave the job and those who do not, and is prone to many misjudgments.

The Naive Bayes model constructs a classifier based on the assumption of feature - conditional independence. However, in the actual employee turnover data, there are often certain correlations between various features, which are difficult to meet the strict independence assumption. The deviation between this assumption and the actual data leads to the Naive Bayes model being unable to fully utilize the correlation information between features when dealing with employee turnover prediction, thus affecting its classification performance. Its AUC value reaches 0.73, which is slightly higher than the KNN model ' s but still falls short of the ideal prediction level.

The Random Forest and Logistic Regression models showed relatively good classification abilities. The Random Forest effectively reduces the overfitting problem that may occur in a single decision tree by integrating multiple decision trees and can explore information in the data from multiple dimensions. In the employee turnover prediction scenario, it can comprehensively consider multiple factors and has a certain ability to capture the complex interaction relationships between different features. Therefore, its AUC value reached 0.78, and to a certain extent, it can more accurately identify employees with a high turnover risk.

As a classic linear classification model, when there is a certain linear relationship between the features related to employee turnover and whether to leave the job, Logistic regression can construct an effective classification model by reasonably assigning weights to these features. In this study, its AUC value is 0.82, indicating that Logistic regression has a certain reliability in employee turnover prediction and can provide prediction results with certain reference value for enterprises.

The LightGBM and CatBoost models performed outstandingly in predicting employee turnover. LightGBM adopts advanced techniques such as the histogram algorithm and Leaf - wise growth strategy, which greatly improves the training efficiency of the model and its ability to process complex data. When facing the complex task of employee turnover prediction, it can quickly extract key information from a large amount of data, accurately identify the important factors affecting employee turnover, and thus achieve more accurate classification predictions, with a high AUC value.

The CatBoost model performed most prominently among many models, with an AUC value as high as 0.92. The CatBoost not only inherits the advantages of the gradient-boosting algorithm but also innovates in handling categorical features and preventing overfitting. In employee turnover prediction, it can make full use of various data collected by enterprises, including features obtained from different management systems and subjective data evaluated through special methods, deeply explore the potential patterns behind the data, and make extremely accurate judgments on whether employees leave the job, demonstrating excellent prediction ability and providing strong support for enterprises to intervene in employee turnover risks in advance[28].

4.3 Influence factor analysis

The CatBoost model was used to predict the turnover risk of front-line workers in the auto-parts manufacturing enterprise. Through the feature importance evaluation mechanism of the model, the influence weights of each feature on the prediction result were quantified, and the top 10 most significant features were screened out[28]. The specific results are shown in the following Table 6.

Table 6 The influence weights of features

Importance Ranking	Feature Name	Feature Proportion (%)
1	Employee's job position	18.7
2	Employee's length of service in this enterprise	16.3
3	Employee's monthly salary	14.5
4	Employee's performance score	12.8
5	Number of times the employee participated in training	11.2
6	Whether introduced through internal referral	9.6
7	Whether in the probation period	8.4
8	Required skill level for operating relevant equipment or completing production tasks	7.5
9	Severity of consequences caused by incorrect operations	6.8
10	Employee's monthly overtime hours	4.2

The above feature proportions are obtained based on the built - in feature importance calculation method of the CatBoost model. By cumulatively calculating the contribution of features at the decision - tree node splitting during the training process, the higher the weight proportion, the greater the role the feature plays in the decision - making process of distinguishing whether employees leave their jobs[29].

5 Empirical analysis of the intervention effect of the catboost model

After completing the training of the CatBoost model and the evaluation of feature importance, this research collaborated with the enterprise's human resources department. According to the model's prediction results, precise targeted interventions were implemented for employees identified as having a high turnover risk, and a systematic one - year tracking and monitoring was carried out. The intervention strategies covered multiple - dimensional measures such as personalized vocational skills training, optimization of the dynamic performance - incentive system, job - suitability adjustment, and customized employee care programs.

The tracking data shows that after the implementation of the intervention measures, the quarterly turnover rate of front- line employees showed a continuous downward trend. A comparative analysis with company turnover rate data in the same period of the last three years indicates that the CatBoost model-based intervention strategy has achieved remarkable

results. Not only did it make the quarterly turnover rate significantly lower than the historical level but also showed stable and obvious improvement effects in different operating cycles of the enterprise, such as the production off-season and peak season(see Fig. 2). The turnover rate shows a consistent downward trend in all quarters after the

implementation of the CatBoost model, with the most significant reduction (over 26%) observed in Q4, the traditional peak turnover period, compared to the same quarter in previous years.

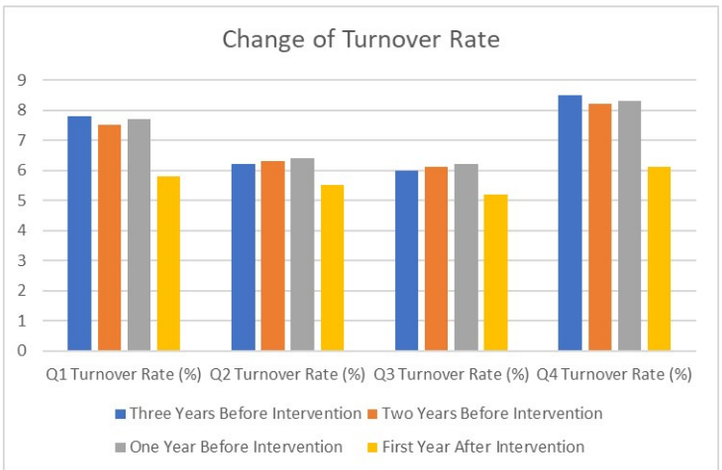

Fig. 2 Change of Turnover Rate

6 Conclusion

This research constructs a CatBoost-based turnover risk prediction model for front- line workers in auto parts manufacturing enterprises using the daily management data of the enterprises. Through comparison with traditional models such as Logistic regression and Random Forest, and verification by one- year intervention tracking, the following core conclusions are drawn:

In terms of model construction and performance evaluation, the CatBoost algorithm shows significant advantages in dealing with turnover risk prediction scenarios with high dimensional data and a large number of categorical features. After hyperparameter tuning, its AUC value on the test set reaches 0.89, and the F1 score is 0.82, which is significantly better than the comparative models, verifying the high efficiency and accuracy of this algorithm in identifying employees with a high turnover risk. The importance analysis of characteristics shows that 10 key characteristics, such as the employee's job position, the length of service, and the monthly salary, have a significant impact on turnover risk, providing a quantitative basis for enterprises to formulate intervention strategies.

At the practical application level, the targeted intervention measures implemented based on the prediction results of the CatBoost model have achieved remarkable results. Through one- year tracking and monitoring, the quarterly turnover rate of the enterprise's front- line employees dropped from about 6.3% initially to 4.5%. In the traditional peak turnover period (the fourth quarter), the turnover rate decreased from an average of 8. 3% in the last three years to 6.1%, a decrease of 26.5%, which is significantly lower than the same period in the last three years. This result indicates that this model can not only accurately predict the risk of turnover, but can also effectively support enterprises in optimizing human resource management strategies, achieving the objectives of reducing the employee turnover rate and improving management efficiency.

This research innovatively applies the CatBoost algorithm to the field of employee turnover prediction in auto parts manufacturing enterprises, providing a scientific and feasible technical solution for the industry. Future research can further explore the applicability of the model in enterprises of different scales and production modes, continuously optimize the model by incorporating more dynamic data (such as employee emotion indicators, industry salary fluctuations, etc.), and extend the intervention tracking period to deeply verify the long- term effectiveness of the model, so as to provide more comprehensive and reliable support for enterprise human resource management decisions.

About the Author

Xiao Xiao, xiaoxiao@gzist.edu.cn.

* Corresponding Author : XueYang Zhong, zxy@gzist.edu.cn.

References

[1] V.Knapić,B.Rusjan, andK.Božič', Importance of first-line employees in lean implementation in smes: a systematic literature review', International Journal of Lean six sigma, vol. 14, no. 2, pp. 277–308, 2023.

[2] M.J. Ligarski, B. Rożałowska, and K. Kalinowski, 'A study of the human factor in industry 4.0 based on the automotive industry', Energies, vol.

14, no. 20, p. 6833, 2021.
[3] Y. Feng et al., 'A study on employee motivation in small and mediumsized automotive parts manufacturing enterprises', Academic Journal of Business Management, vol. 7, no. 2, pp. 1–7, 2025.
[4] G. Mao, B. Hu, and H. Song, 'Exploring talent flow in wuhan automotive industry cluster at china', International Journal of Production Economics, vol. 122, no. 1, pp. 395–402, 2009.
[5] K. Moon, P. Loyalka, P. Bergemann, and J. Cohen, 'The hidden cost of worker turnover: Attributing product reliability to the turnover of factory workers', Management science, vol. 68, no. 5, pp. 3755–3767, 2022.
[6] M. Climek, R. Henry, and S. Jeong, 'Integrative literature review on employee turnover antecedents across different generations: commonalities and uniqueness', European Journal of Training and Development, vol. 48, no. 1/2, pp. 112–132, 2024.
[7] M. Lazzari, J. M. Alvarez, and S. Ruggieri, 'Predicting and explaining employee turnover intention', International Journal of Data Science and Analytics, vol. 14, no. 3, pp. 279–292, 2022.
[8] S.-Y. Chen, A. Y.-P. Lee, and D. Ahlstrom, 'Strategic talent management systems and Employee behaviors: the mediating effect of calling', Asia Pacific Journal of Human Resources, vol. 59, no. 1, pp. 84–108, 2021.
[9] M. Al Akasheh, E. F. Malik, O. Hujran, and N. Zaki, 'A decade of research on machine learning techniques for predicting employee turnover: A systematic literature review', Expert Systems with Applications,vol.238, p. 121794, 2024.
[10] N. Jain, A. Tomar, and P. K. Jana, 'A novel scheme for employee churn problem using multi-attribute decision making approach and machine learning', Journal of Intelligent Information Systems, vol. 56, pp. 279–302, 2021.
[11] Y. Yan and S. Huimin, 'Empirical research on the influencing factors of employee turnover of private auto parts enterprises', in 2009 International Conference on Information Management, Innovation Management and Industrial Engineering, vol. 4. IEEE, 2009, pp. 131–135.
[12] S.Ponmalar, N.A.Fowmiya,andC.Nandhini',Ai-driven retention: A hybrid approach to employee turnover prediction', in 2024 9th International Conference on Communication and Electronics Systems (ICCES). IEEE, 2024, pp. 1554–1559.
[13] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, 'Catboost: unbiased boosting with categorical features', Advances in neural information processing systems, vol. 31, 2018.
[14] J. Hancock and T. Khoshgoftaar, 'Catboost for big data: an interdisciplinary review', Journal of big data 7 (1): 94, 2020.
[15] A. A. Ibrahim, R. L. Ridwan, M. M. Muhammed, R. O. Abdulaziz, and G. A. Saheed, 'Comparison of the catboost classifier with other machine learning methods', International Journal of Advanced Computer Science and Applications, vol. 11, no. 11, pp. 738–748, 2020.
[16] C. Kulkarni, 'Advancing gradient boosting: A comprehensive evaluation of the catboost algorithm for predictive modeling', Journal of Artificial Intelligence, Machine Learning and Data Science, vol. 1, no. 5, pp. 54–57, 2022.
[17] L. Zhang and D. Jánošík, 'Enhanced short-term load forecasting with hybrid machine learning models: Catboost and xgboost approaches', Expert Systems with Applications, vol. 241, p. 122686, 2024.
[18] J. Hancock and T. M. Khoshgoftaar, 'Medicare fraud detection using catboost', in 2020 IEEE 21st international conference on information reuse and integration for data science (IRI). IEEE, 2020, pp. 97–103.
[19] D. Avrahami, D. Pessach, G. Singer, and H. Chalutz Ben-Gal, 'A human resources analytics and machine-learning examination of turnover: implications for theory and practice', International Journal of Manpower, vol. 43, no. 6, pp. 1405–1424, 2022.
[20] J. Park, Y. Feng, and S.-P. Jeong, 'Developing an advanced prediction model for new employee turnover intention utilizing machine learning techniques', Scientific Reports, vol. 14, no. 1, p. 1221, 2024.
[21] R. Chakraborty, K. Mridha, R. N. Shaw, and A. Ghosh, 'Study and prediction analysis of the employee turnover using machine learning approaches', in 2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON). IEEE, 2021, pp. 1–6.
[22] H. Mulang, 'Analysis of the effect of organizational justice, worklife balance on employee engagement and turnover intention', Golden Ratio of Human Resource Management, vol. 2, no. 2, pp. 86–97, 2022.
[23] K. Maharana, S. Mondal, and B. Nemade, 'A review: Data pre-processing and data augmentation techniques', Global Transitions Proceedings, vol. 3, no. 1, pp. 91–99, 2022.
[24] K. Cho, T. Yoon, S. Park, and D. Park, 'Using standardization for fair data evaluation', in 2018 20th International Conference on Advanced Communication Technology (ICACT). IEEE, 2018, pp. 585–589.
[25] V.C.Nitesh, 'Smote: synthetic minority over-sampling technique', J Artif Intell Res, vol. 16, no. 1, p. 321, 2002.
[26] S. Wang, Y. Dai, J. Shen, and J. Xuan, 'Research on expansion and classification of imbalanced data based on smote algorithm', Scientific reports, vol. 11, no. 1, p. 24039, 2021.
[27] P. Schratz, J. Muenchow, E. Iturritxa, J. Richter, and A. Brenning, 'Hyper-parameter tuning and performance assessment of statistical and machine learning algorithms using spatial data', Ecological Modelling, vol. 406, pp. 109–120, 2019.
[28] W. Patrick Walters, 'Comparing classification models - a practical tutorial', Journal of Computer-Aided Molecular Design, vol. 36, no. 5, pp. 381–389, 2022.
[29] F. Zhang, J. Yang, B. Liang, and H. He, 'Analysis of influencing factors of new energy vehicle satisfaction based on scenario thinking and catboost model', in IOP Conference Series: Earth and Environmental Science, vol. 769, no. 4. IOP Publishing, 2021, p. 042023.