2025 5th International Conference on Artificial Intelligence and Industrial Technology Applications (AIITA)

Preprediction Method of Enterprise Migration Behavior Based on XGBoost

The First author: Nan Wu
China Academy of Industrial Internet
Chaoyang,Beijing,China
Email: wunan@china-aii.com

The Second author: Anqi Zhao
China Academy of Industrial Internet
Chaoyang,Beijing,China
Email: zhaoanqi@china-aii.com

The Third author: Jiehao Chen
China Academy of Industrial Internet
Chaoyang,Beijing,China
*Corresponding author: chenjiehao@china-aii.com

The Fouth author: Haowei Li
China Academy of Industrial Internet
Chaoyang,Beijing,China
Email: lihaowei@china-aii.com

***Abstract*—The paper adopts the method of data mining and machine learning, through the data mining, analyzes the key factors affecting the enterprise migration, the main factors including the policy environment, geographical location, enterprise development rules, competitive status, and using XGBoost algorithm to model and predict the enterprise migration behavior. The results show that XGBoost algorithm has high accuracy and practicability in the prediction of migration behavior, which can provide strong support for the decision-making of relevant government departments, and can significantly improve the work efficiency of relevant personnel.**

Keywords-XGBoost; machine learning; enterprise migration; predictive model

I. INTRODUCTION

In the face of the increasingly complex international situation, China attaches great importance to the construction of the resilience and safety level of the industrial chain and supply chain, and proposes the construction of a national important industrial backup base. How to guide enterprises in the industrial chain to carry out orderly transfer in China and carry out scientific investment promotion is a necessary topic for governments at all levels. At present, local governments mainly rely on traditional models and channels, mainly on relational investment, opportunity-oriented investment and wide-net investment. There are common problems such as waste of resources, low conversion rate of investment, short regional industrial chain and poor effect of industrial agglomeration. It is practically urgent to use technical means to build the model and to intelligently recommend high-quality enterprises with the willingness to migrate, so as to make targeted investment contact.

This paper selects machine learning model with accuracy and calculation efficiency of XGBoost model, put forward the XGBoost based grid supplier performance risk prediction model, combined with the actual business analysis, first through the feature engineering structure 87 features for initial training, through model optimization, selected 26 features training again, through the model output into specific probability, can further guide the practical application, assist related business personnel decision. The contributions and innovations of this article mainly include:

1)In view of the current lack of prediction of enterprise migration behavior in the process of attracting investment, relevant data are deeply mined, and characteristic projects are established to further form a complete data set of enterprise migration behavior.

2) The XGBoost-based enterprise migration behavior prediction model was established, and the effectiveness and accuracy of the model were verified through six processes, including data cleaning, feature engineering, model training, optimization, evaluation and application.

II. METHOD

A. Data cleaning

Before processing the above raw data, data cleaning is first performed to improve the accuracy of the model. The data cleaning method mainly include missing value and outlier handling.

In due to business reasons, part of the original data in the problem of missing data, need relevant value processing missing fields, for category of variables we regard missing value as a characteristic value for processing, and for continuous variables we generally use mean, median replacement or use random forest according to the random tree to evaluate the probability of each tree to estimate[4].

The abnormal data generated in the original data collection, processing and transmission process is easy to affect the robustness of the model, so the outliers need to be handled. In this paper, iforest isolated forest is mainly used to identify the outliers and replace the mean value.

B. Feature extraction, construction, and feature selection

Characteristic engineering is the core of the enterprise migration behavior prediction work, which directly determines the quality and application effect of the model. The specific implementation method is as follows:

1) Deeply understand the business and sort out the influencing factors of enterprise migration.

Through interviews with experts in the field and questionnaire surveys, the reasons for enterprise migration are sorted out. Impact the cause of the industrial chain enterprise

979-8-3315-0976-7/25/$31.00 ©2025 IEEE

migration analysis summarized as follows: one is the policy to encourage related industrial transfer lead to relocation demand, in 2018, the Ministry of Industry and Information Technology to revise the previous documents, released the industrial transfer guidance directory (2018), local governments have put forward "vacate cage in bird" "out of the city into the garden" industrial policy, prompting enterprises to migrate. Second, the needs generated by the development of the enterprises own business, such as the new production line, project incubation registration, and the needs of new enterprises generated by the diversified development of enterprises, to promote the enterprises to migrate and seek the most favorable development location. Third, follow the migration caused by the adjustment of the layout of the core enterprises in the industrial chain. For example, in 2008, Hefei, Anhui Province successfully introduced BOE, and in the following years, nearly 1,000 related supporting enterprises moved to Hefei with the company.

2) Preliminary construction of features through exploratory data analysis (EDA), feature combination and other methods.

Exploratory data analysis is the use of statistical and visual methods to explore the data and identify potentially useful features. For example, through the exploratory data analysis of enterprise investment and financing events, it can be found that region is an important factor affecting enterprise migration, and many enterprises will give priority to migration in the province; nearly a quarter of the sample enterprises have cross-industry investment behavior; and about 21% will invest again. Based on the above rules, we can initially construct the relevant index characteristics. In addition, based on the above features, these features can be combined to create new features, such as multiplication and division between different features.

3) Feature selection is realized by using filtering method and embedding method.

Through the methods mentioned above, we have preliminarily constructed some feature indicators. The selection of necessary features mainly employs two approaches: one is the use of filtering for feature selection. This paper adopts variance screening. Features with higher variances can be considered more useful. If the variance is low, such as less than 1, then this feature may not significantly impact our algorithm. In the extreme case, if a feature has a variance of 0, meaning all samples have the same value for that feature, it will have no effect on model training and can be discarded directly. Additionally, the chi-square test is used to examine the correlation between a features distribution and the output value distribution. This is primarily done using the chi-square class in sklearn API within Python, obtaining the chi-square values and significance levels P for all features. By setting a threshold for the chi-square value, features with larger chi-square values are selected; the second approach is the use of embedding for feature selection. It involves training certain machine learning algorithms and models to obtain the weight coefficients of each feature, then selecting features based on their weight coefficients from highest to lowest. Similar to the filtering method, Similar to filtering, but it determines the quality of features through machine learning training, rather than directly determining the quality of features from some statistical indicators.Here we call sklearn API and use the SelectFromModel function to select features. Finally, we screened 26 indicators from 87 indicators, as shown in TABLE 1.

TABLE 1. FINAL INDICATORS OF THE ENTERPRISE MIGRATION BEHAVIOR PREDICTION MODEL

order number	name of index	order number	name of index
1	Whether there are cross-industry enterprises in recent N years	14	Whether there are key remote core enterprises
2	Whether there is a cross-industry investment in recent N years	15	Spatial distance with the core enterprises
3	Whether there are key core enterprises	16	Enterprise employee size range
4	Whether the existing entity and the core enterprise are in the same region	17	Whether the enterprise is in the industrial cluster
5	The number of key trading targets coincides with the companys region	18	The same industry enterprise relocation activity
6	Whether an existing entity is in the same area with the key objects	19	Matching degree with the local industrial policies
7	Enterprise sales growth rate	20	Enterprise scale classification
8	Industry financing activity	21	Rate of change of core product yield
9	enterprise age	22	The GDP growth rate of the cities where the enterprises are located
10	There are no different to set up enterprises	23	market share
11	Corporate financing activity	24	registered capital
12	Enterprise sales scale	25	Enterprise nature
13	Total amount of enterprise patent applications	26	Whether the company has a branch office

C. XGBoost model

1)loss function of the XGBoost

$$L_t = \sum_i^m L(y_i, f_{t-1}(x_i) + h_t(x_i)) + \gamma J + \frac{\lambda}{2}\sum_{j=1}^J w_{tj}^2 \quad (1)$$

Finally, we want to minimize the above loss function and get the optimal solution w_{tj} for all J leaf node regions and the decision tree. XGBoost Instead of fitting the first derivative of the Taylor expansion as in GBDT, it is expected to be solved directly based on the second order Taylor expansion of the loss function[1,7]. Now we look at the second-order Taylor expansion of this loss function:

$$L_t = \sum_i^m L(y_i, f_{t-1}(x_i) + h_t(x_i)) + \gamma J + \frac{\lambda}{2}\sum_{j=1}^J w_{tj}^2$$

$$\approx \sum_{i=1}^m (L(y_i, f_{t-1}(x_i)) + \frac{\partial L(y_i, f_{t-1}(x_i))}{\partial f_{t-1}(x_i)} h_t(x_i) + \frac{1}{2}\frac{\partial^2 L(y_i, f_{t-1}(x_i))}{\partial^2 f_{t-1}(x_i)} h_t^2(x_i) + \gamma J + \frac{\lambda}{2}\sum_{j=1}^J w_{tj}^2 \quad (2)$$

For convenience, we record the first and second derivatives of the weak learner as:

$$g_{ti} = \frac{\partial \mathrm{L}(\mathrm{y_i},\mathrm{f_{t-1}}(\mathrm{x_i}))}{\partial \mathrm{f_{t-1}}(\mathrm{x_i})}, h_{ti} = \frac{\partial^2 \mathrm{L}(\mathrm{y_i},\mathrm{f_{t-1}}(\mathrm{x_i}))}{\partial^2 \mathrm{f_{t-1}}(\mathrm{x_i})} \quad (3)$$

$then\ our\ loss\ function\ can\ now\ be\ expressed\ as$:

$$L_t \approx \sum_{\mathrm{i=1}}^{\mathrm{m}}(\mathrm{L}(\mathrm{y_i},\mathrm{f_{t-1}}(\mathrm{x_i}))\ +\ g_{ti}\mathrm{h_t}(\mathrm{x_i}) + \frac{1}{2}h_{ti}(\mathrm{x_i}) + \gamma J\ + \frac{\lambda}{2}\sum_{j=1}^{J} w_{tj}^2 \quad (4)$$

In the loss function,$\mathrm{L}(\mathrm{y_i},\mathrm{f_{t-1}}(\mathrm{x_i}))$is a constant, which has no effect on the minimization and can be removed. Meanwhile, since the value of the j the leaf node of each decision tree will eventually be the same value wt j, our loss function can continue to degenerate.

$$\begin{aligned} L_t &\approx \sum_{\mathrm{i=1}}^{\mathrm{m}}(\mathrm{L}(\mathrm{y_i},\mathrm{f_{t-1}}(\mathrm{x_i}))\ +\ g_{ti}\mathrm{h_t}(\mathrm{x_i}) + \frac{1}{2}h_{ti}(\mathrm{x_i}) + \gamma J\ + \frac{\lambda}{2}\sum_{j=1}^{J} w_{tj}^2 \\ &= \sum_{j=1}^{J}(\sum_{\mathrm{x_i \in R_{ij}}} g_{ti}\, w_{tj}) + \frac{1}{2}\sum_{\mathrm{x_i \in R_{ij}}} h_{ti}\, w_{tj}^2) + \gamma J\ + \frac{\lambda}{2}\sum_{j=1}^{J} w_{tj}^2 \\ &= \sum_{j=1}^{J}[(\sum_{\mathrm{x_i \in R_{ij}}} g_{ti})w_{tj} + \frac{1}{2}(\sum_{\mathrm{x_i \in R_{ij}}} h_{ti} + \gamma)w_{tj}^2] + \gamma J \quad (5) \end{aligned}$$

We separately represent the sum of the first and second derivatives of each leaf node region sample as follows:

$$\mathrm{G_{tj}} = \sum_{\mathrm{x_i \in R_{ij}}} g_{ti}\, , H_{tj} = \sum_{\mathrm{x_i \in R_{ij}}} h_{ti} \quad (6)$$

The final loss function can be expressed as:

$$L_t = \sum_{j=1}^{J}[\mathrm{G_{tj}} + c(H_{tj} + \gamma)w_{tj}^2] + \gamma J \quad (7)$$

2) the loss function of XGBoost

Take the $\mathrm{w_{tj}}$ based on the loss function and make the derivative 0. In this way, the optimal solution $\mathrm{w_{tj}}$ expression of the leaf node region is:

$$w_{tj} = -\frac{\mathrm{G_{tj}}}{H_{tj}+\lambda} \quad (8)$$

When w_{tj} takes the optimal solution, the expression corresponding to the original loss function is:

$$L_t = \sum_{j=1}^{J}\frac{G_{ti}^2}{H_{tj}+\lambda} + \gamma J \quad (9)$$

It is best to minimize the loss of the loss function every time we split the left and right subtrees. That is, assuming that the first order second derivative sum of the left and right subtrees of the current nodes are GL, HL, GR, HL, then we expect to maximize the following equation:

$$\max\frac{1}{2}\frac{G_L^2}{H_L+\lambda} + \frac{1}{2}\frac{G_R^2}{H_R+\lambda} - \frac{1}{2}\frac{(G_L+G_R)^2}{H_L+H_R+\lambda} - \gamma \quad (10)$$

III. Experimental Results And Analysis

A. Data set

This paper modeling the overall data for 50000 enterprises, including negative sample enterprise 46000, positive sample enterprise 4000 (positive and negative sample definition and screening rules see TABLE 2), data dimensions including enterprise management data, industrial and commercial data, macroeconomic data, intellectual property data, investment and financing data and statistical data in the regional fundamental data, etc.Through characteristic engineering, 87 risk characteristics affecting the performance of suppliers are output, and then the experimental data set of this paper is obtained. The data set was partitioned using the train _test_split function of the sklearn package using the default test_size parameter =0.25, i. e., the training set 75% and the test set 25%[2,3].

TABLE 2. DEFINITION OF POSITIVE AND NEGATIVE SAMPLES AND SAMPLE SCREENING RULES

class	description
Definition of positive and negative samples	Negative sample: the enterprises that did not initiate the migration behavior. Positive sample: the enterprise with the migration behavior.
Sample screening rules	（1）Unified social credit code is as follows: 91 beginning, which is for industrial and commercial enterprises (excluding individual industrial and commercial households, non-profit social public welfare organizations, etc.) （2）The elimination of cancellation, cancellation, revocation and other enterprises can indicate that the enterprises has stopped normal operation, and the enterprises still existing. （3）By comparing the change information of the registered address of the enterprise, it is judged whether the enterprise has migrated. Before the enterprise is marked as 0 (negative sample), the mark is marked as 1 (positive sample), and the mark that cannot be judged due to the lack of field information is 2.

B. Model parameters and model validation

1)Model parameter tuning method

Using Grid Search, Random Search,Bayesian Optimization and other methods, all hyperparameters of XGBoost are optimized, and the final model parameters are shown in TABLE 3.

a) Grid Search: Grid search sets the possible values of hyperparameters into a grid and comprehensively traverses this grid, using cross-validation to evaluate the performance of each hyperparameter combination, ultimately selecting the best-performing combination[5]. The main workflow includes defining the hyperparameter space, constructing the grid, cross-validation, evaluation, and selection. We use GridSearchCV from sklearn to implement this.

b) Random Search: Random search is an optimization method for hyperparameters. It involves randomly sampling multiple combinations of hyperparameters within a predefined hyperparameter space, training and evaluating each combination to find the best-performing set[6]. Unlike grid search, random search does not exhaust all possible combinations but randomly selects some for evaluation. We implement this using RandomizedSearchCV from sklearn.

c) Bayesian Optimization: Bayesian optimization is an intelligent hyperparameter tuning method that constructs a surrogate model to approximate the objective function and selects the optimal hyperparameter combination based on this surrogate model. Specifically, Bayesian optimization uses Gaussian processes or other regression models as surrogate models, gradually exploring and utilizing information from the objective function to find the optimal solution. We implement this using the hyperopt library[8].

TABLE 3. THE XGBOOSTT MODEL PARAMETERS

order number	The parameter name	parameter values
1	n_estimators	370
2	max_depth	6
3	eval_metric	'auc '
4	min_child_weight	1
5	subsample	0.5
6	colsample_bytree	0.8
7	colsample_bylevel	0.9
8	learning_rate	0.1
9	gamma	0.9
10	reg_alpha	1
11	reg_lambda	1
12	reg_lambda	1

2)Model verification method

We used auc, precision, accuracy, f1score, recall and so on to verify the model and save the trained model locally,as shown in FIGURE 1 and TABLE 4.

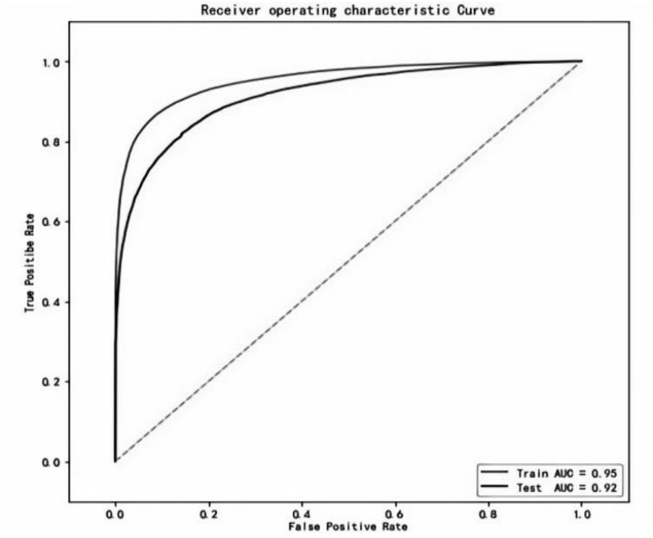

FIGURE 1. MODEL AUC DIAGRAM (XGBOOST)

TABLE 4. XGBOOST MODEL VALIDATION

	test set	training set
auc	0.92	0.95
precision	0.87	0.89
accuracy	0.83	0.89
f1 score	0.81	0.88
recall	0.74	0.85

C. Comparison of the experimental results of multiple models

The above experimental procedure is based on the XGBoost algorithm. In order to further verify the XGBoost compared to other machine learning model in the supplier of performance risk superiority and availability, select seven kinds of mainstream models in the field of machine learning: SVM, decision tree, KNN, random forest, XGBoost, neural network, NB, using the same training set training, test set test, compare the experimental results as shown in TABLE 5. Based on the XGBoost model score is better than the other six algorithms, fully showing the performance advantage of XGBoost.

TABLE 5 . COMPARISON OF SEVERAL CLASSIFICATION ALGORITHMS

Algorithm name	score
XGBoost	0.86
neural network	0.84
random forest	0.84
svm	0.81
knn	0.80
decision tree	0.67
NB	0.55

D. Experimental results

The trained model can be used to predict the migration behavior of enterprises. Here, we predict the possibility of enterprises to Changsha city, and obtain the probability of the migration of relevant enterprises.The results are shown in TABLE 6.

TABLE 6 . APPLICATION OF PREDICTING ENTERPRISE MIGRATION BEHAVIOR (PART)

order number	corporate name	migration probability
1	Beijing Chang so Culture Media Co., LTD	0.798964416
2	Beijing Jinyiheng Construction Engineering Co., Ltd	0.014510836
3	Beijing Hehe Medical Diagnostic Technology Co., Ltd	0.91132457
4	Beijing Taihao Intelligent Engineering Co., LTD	0.914515848
5	Efa Energy Engineering Co., Ltd	0.651987258
6	Tianjin Proficient Control Instrument Technology Co., LTD	0.931951152
7	Deyou (Tianjin) Real Estate Brokerage Service Co., Ltd	0.907618517
8	Tianjin Tiens Biological Engineering Co., Ltd	0.949873589
9	Tianjin Dean Construction Engineering Co., Ltd	0.506201239
10	Tianjin Supai Automobile Trading Co., LTD	0.861988465
11	Changsha Lola Fast Running Technology Partnership (Limited partnership)	0.550403318

Model prediction is mainly programmed by Python, specifically implemented as follows:

1) Read the new data samples.

2) Process the data through the data preprocessing module program code.(Including missing data value handling, outlier cleaning, etc.)

3) Call the locally saved trained model to complete the model prediction.

IV. Conclusion

This paper uses XGBoost algorithm to construct the prediction model of enterprise migration behavior, which aims to provide scientific basis for government investment attraction and enterprise decision-making. The results show that the XGBoost algorithm has high accuracy and utility in predicting enterprise migration behavior. Through in-depth mining of enterprise migration behavior data and feature engineering, we successfully constructed a dataset containing 26 key indicators, and used these features to train an efficient prediction model. The experimental results verify the effectiveness of the model, which is excellent in AUC, accuracy, accuracy and F1 score, and is better than other traditional machine learning models. Moreover, the practical application prediction results of the model also show its potential in predicting enterprise migration behavior. In conclusion, this study not only provides a new perspective for understanding enterprise migration behavior, but also provides a powerful tool for ecision-makers in related fields.

References

[1] Budholiya K ,Shrivastava S K,Sharma V.An optimized XGBoost based diagnostic system for ellective prediction of heartdisease[J].Journal of King Saud University-Computer and Information Seiences , 2022, 34: 4514-4523

[2] Al Ali A , Khedr Ahmed M M,El-Bannany M,et al.A powerful predicting model for financial statement fraud based onoptimized XGBoost ensemble learning technique[J].Applied Sciences-Basel, 2023, 13 (4) :2272

[3] Liu Jiang,Xu Kangzhi, Cai Baigen ,et al. Fault prediction of on-board train control equipment using a CGAN-enhanced XGBoost method with unbalanced samples[J].Machines ,2023,11(1):114

[4] REN M J,JIN G Q,WANG X w,et al. Microblog Popularity Prediction Algorithm Based on XGBoost [J]. Data Acquisitionand Processing, 2022, 37(2):383-395.

[5] ZHUANG J Y,YANG G H,ZHENG H F, et al. CNN-LSTM XGBoost short-term power load forecasting method based onmulti-model fusion[J]. Electric Power,2021,54(5):46-55.

[6] ZHU JX,ZOU X S, XIONG w, et al. Short-Term Power Load Forecasting Based on Prophet and XGBoost Mixed Model [J].Modern Electric Power,2021,38(3):325-331.

[7] LEI X N,LIN L F,XIAO B Q, et al. Reexploration of default characteristics of small and micro enterprises: machine learning model based on SHAP interpretation method[J]. Chinese Jour-nal of Management Science, 2021,27:1-13.

[8] LIAO B,WANG Z N,Ll M,et al. Integrating XGBoost and SHAP model for football player value prediction and characteri-stic analysis[J]. Computer Science,2022,49(12):195-204.