J. COMBIN. MATH. COMBIN. COMPUT. 127b (2025) 341--357

Journal of Combinatorial Mathematics and Combinatorial Computing
www.combinatorialpress.com/jcmcc

Using Logistic Regression Modeling to Analyze Barriers and Facilitators to Social Integration of Foreign Immigrants

Wenjie Huang[1, ✉]

[1] *Department of Public Security, Jilin Police College, Changchun, Jilin, 130117, China*

ABSTRACT

Due to differences in lifestyle, cultural capital and social support, foreign immigrants often have difficulty integrating into the ecology of their native communities and are limited in their space for development. To solve this difficulty, this paper applies the principle of regularization to obtain a logistic regression model by categorizing the factors affecting the social integration of foreign immigrants. The algorithms of log-likelihood function and negative Hessian matrix are used to optimize the parameters of the model, construct the multivariate logistic regression model based on the social integration of foreign immigrants, and analyze the regression results among various factors. The success rate of foreign immigrants' local integration is higher when the immigration-related system is more perfect, the foreign immigrants' cultural identification with the local area is higher, the cognitive deviation between foreign immigrants and locals is smaller, and the community integration structure is more appropriate. The highest correlation between the factors affecting the social integration of foreign immigrants is the formation of ethnic networks that are not embedded in the community by foreign immigrants who “embrace the group”, and the cognitive bias of local residents towards foreign immigrants, with a correlation coefficient of 0.9214, and the correlation coefficients of the rest of the indicators are less than 0.9. This paper classifies the migrants into “migrants of work nature” and “migrants of employment nature” in accordance with the purpose of their migratory activities. In this paper, according to the purpose of migration, migrants are classified into four categories: “work migration, study migration, investment migration and shelter migration”, and the results of the multivariate logistic regression analysis are credible.

Keywords: foreign immigrants, regularization principle, multiple logistic regression model, log-likelihood function, negative Hessian matrix

✉ Corresponding author.
E-mail address: huangwenjie0331@163.com (W. Huang).

Received 05 January 2024; Revised 15 April 2024; Accepted 20 August 2024; Published Online 16 April 2025.
DOI: 10.61091/jcmcc127b-021

1. Introduction

Since the 21st century, with the gradual deepening of global integration, the interaction between countries, groups and individuals has become more and more frequent, and the wave of immigrants has swept across the globe, with China also playing an important role in the wave of international immigrants. At the same time when foreign immigrants flood into Chinese cities to settle down, internationalized communities, which are the "testing ground" for Chinese and foreign residents to build and govern together, have also gradually become the focus of social attention [1-4]. Transnational immigrants come to China for many purposes, most of them come to China to seek better development because of their work, and the social integration of international immigrants and domestic immigrants has slowly come into the attention of academics [5-6]. In recent years, Beijing, Shanghai, Shenzhen, Qingdao and other major cities in China have launched internationalized community construction planning goals, actively promoting the strategy of urban internationalization, and expanding the level of opening up to the outside world with a view to enhancing the competitiveness of international talents [7-8].

With the continuous development of the market economy, the degree of openness of Chinese cities is getting higher and higher, the attraction of Chinese cities to foreigners is increasing, and the number of international immigrants in Chinese cities is growing, these people work and live in Chinese cities for a long time, and even set up their families locally in Chinese cities, and their sense of belonging to Chinese cities is getting stronger and stronger, and the opportunities for development in Chinese cities are also growing significantly [9-10]. It is undeniable that the agglomeration of foreign immigrants brings diverse colors and vitality to internationalized communities, but it also brings new challenges to urban community governance. Foreign immigrants often have a strong psychological identification with their home country's culture and original identity, a strong sense of alienation from their communities, and a lower degree of community participation; due to differences in lifestyle, cultural capital and social support, conflicts and disputes between foreign and local residents in the community occur frequently [11-13]. All these integration problems not only weaken the willingness of immigrants, especially high-level talents, to stay in the place of migration, but also prevent them from giving full play to their role in promoting the construction and development of the local city; it also makes the immigrant groups more dependent on intra-ethnic relations, forming an ethnic network that is not embedded in the national governance system, and this governance "vacuum" may be alienated to the immigrant communities for drug use, drug abuse, and other social activities. This governance "vacuum" may be alienated into a breeding ground for migrants to engage in illegal and criminal activities such as drug abuse, fraud, and illegal missionary activities, threatening national security and social order [14-16]. In this context, exploring the social integration process of transnational populations is of great significance in promoting exchanges between countries and building a healthy, orderly and harmonious immigrant community.

Generally speaking, China's foreign immigrant management policy has made great progress from strict control to both management and service, but there is still a scarcity of "attention" in foreign immigrant integration and service guarantee [17]. How to effectively link the immigrant integration system with the concept of national security and closely integrate it with the strategy of strengthening the country with talents, to break through the realities of the integration of foreign immigrants in the internationalized community, to guide them to understand and accept China's mainstream values, institutional system and cultural traditions, to improve their willingness to abide by the law and to restrain their misconduct, is the core of the problem that should be paid attention to [18-19].

In this paper, four aspects of barrier factors affecting the social integration of foreign immigrants are organized through a survey. After that, a logistic regression classification model is introduced to optimize the overfitting phenomenon of the regression model by introducing the principle of regularization. Considering the diversity of immigrant groups, this paper promotes the binary logistic regression model to get the form of multivariate logistic regression model. Three important function conditions for the parameter vector of "log-likelihood function, score function and negative Hessian matrix" are obtained. Finally, the multivariate logistic regression model of foreign immigrants' social integration is established, and the barriers and facilitating factors of foreign immigrants' social integration are analyzed.

2. Multiple logistic regression analysis model construction

2.1. Obstacles to the integration of foreign migrants

2.1.1. Absence of a system for the integration of migrants. The legal and policy system for immigration is becoming more and more developed, but there is insufficient supply of systems related to the integration of foreign immigrants: at the macro level, the law on the integration of immigrants is not in place; at the meso level, the policy on services and guarantees for foreign immigrants is not embedded; and at the micro level, the evaluation mechanism for the integration of foreign immigrants in internationalized communities is not in place, which leads to a lack of rules and regulations for the internationalized communities to follow in their practice of facilitating the integration of foreign immigrants.

2.1.2. Structural imbalance in community integration. At present, the practice of foreign immigrant integration in internationalized communities does not seem to have evolved in full accordance with the logic of the balance between state embeddedness and social autonomy; on the one hand, the administrative power is not sufficiently strong in embedding immigrant integration practices, and on the other hand, the lack of vitality in the participation of foreign immigrants and social organizations in the community, which leads to an imbalance in the structure of integration in the community.

2.1.3. Lack of cultural identity. Cultural identity, on the premise of respecting cultural plurality, community members strengthen the sharing of information and the exchange of cultural values through participation in certain forms of cultural activities, so as to reach a consensus and recognition of culture. The same cultural identity is the guarantee of community group cohesion. Diverse foreign immigrants gather in internationalized communities, which also brings about multicultural clustering. However, there is no opportunity for exchanges and integration between multiple cultures, and they remain relatively independent and closed. The break in the cultural field and the de-mosaicing of cultural identities have led to the mosaicization of the overall cultural space of the internationalized community.

2.1.4. The "immigrant-native" two-way cognitive bias. As integration subjects, foreign immigrants are accustomed to relying on intra-ethnic networks to rebuild social support, and are in a state of subjective disorientation in the integration structure, with little subjective willingness to integrate; while local residents of the community, influenced by the stereotype that immigrants take up employment opportunities, perceive foreign immigrants as resource-consuming populations rather than as positive elements for economic development, and in practice do not have a high degree of acceptance of foreign immigrants.

2.2. Logistic regression model

2.2.1. Introduction to the model. Logistic regression [20] models are usually applied to binary classification tasks. That is, in the logistic regression model for binary classification, for each independent variable x, the corresponding y value has only two value cases, i.e., {0, 1}. The logistic regression model can be well solved for the binary classification task with good performance.

Therefore, the logistic regression model is used in this paper as a benchmark model for predicting the social integration of foreign immigrants.

Classification is the most common class of machine learning algorithms. Logistic regression belongs to supervised learning, so it is necessary to collect a batch of labeled and processed data as a training set in advance, which consists of a set of training data with features and labels of the data:

$$D = (x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n) \tag{1}$$

where x_i is a m -dimensional vector and y_i takes values in {0, 1}. Thus, this classification task can be reduced to the question of how to find a decision function $y^* = f(x)$ such that it can perform well enough on an unknown dataset.

Before introducing the logistic regression model, a function, the sigmoid function [21], also called the kernel function, is introduced in the mathematical form:

$$g(z) = \frac{1}{1+e^{-z}} \tag{2}$$

And the corresponding function curve of the kernel function can be seen that this function is an S-shaped curve, which takes values between [0, 1], and the value of the function will be very close to 0 or 1 very quickly away from 0, and $g(0) = 0.5$.

The logistic regression model [22] can be represented by the conditional probability distribution $P(y \mid x;\theta)$, whose mathematical expression for the conditional probability is:

$$P(y = 1 \mid x;\theta) = g(\theta^T x) = \frac{1}{1+e^{-\theta^T x}} \tag{3}$$

$$P(y = 0 \mid x;\theta) = 1 - P(y = 1 \mid x;\theta) = \frac{1}{1+e^{\theta^T x}} \tag{4}$$

$$\theta^T x = \theta_0 + \theta_1 x_1 + \cdots + \theta_n x_n \tag{5}$$

The corresponding decision function is: $y^* = 1, \text{if } P(y = 1 \mid x;\theta) > 0.5$

where x is the input of the model, corresponding to the feature vector of the factors influencing the integration of foreign immigrants in the dataset; $y \in \{0,1\}$ is the output of the model, corresponding to the prediction results; and θ is the vector of the feature weights, i.e., the parameters of the desired solution. The selection of 0.5 as the threshold is not fixed, of course, different thresholds can be selected according to different situations in practical applications. For a given sample data, if the value of $P(y = 1 \mid x;0)$ is greater than 0.5, the sample is categorized as a positive case.

In the logistic regression model, for a given input x, the corresponding output y is given by its decision function, there is a high probability that there will be an error between the predicted value of this output and the true labeled value, inch is a loss function can be used to pick a measure of the degree of prediction error, the smaller the value of the loss function, it indicates that the error between the predicted value and the true labeled value is smaller, i.e., the more the model's performance is more and more excellent. The loss function of the logistic regression model is shown in the following equation, i.e.:

$$Cost(g(\theta^T x), y) = \begin{cases} -\log(g(\theta^T x)) \\ -\log(1 - g(\theta^T x)) \end{cases} \text{if} \begin{cases} y = 1 \\ y = 0 \end{cases} \tag{6}$$

Since the value of y can only be 0 or 1, the above equation can be combined, i.e.:

$$Cost(g(\theta^T x), y) = -y\log(g(\theta^T x)) - (1-y)\log(1 - g(\theta^T x)) \tag{7}$$

For logistic regression models, maximum likelihood estimation can be used to solve for the parameters, i.e., to find a set of parameters under which the value of the likelihood function for the training data is maximized. The likelihood function in a logistic regression model can be expressed as:

$$L(\theta)=\prod_{i=1}^{m}P(y_i\mid x_i;\theta)=\prod_{i=1}^{m}g(\theta^T x_i)^{y_i}(1-g(\theta^T x_i))^{1-y_i} \tag{8}$$

Taking the logarithm of the above equation yields the following form, called the log-likelihood function:

$$l(\theta)=\sum_{i=1}^{m}\left[y_i\log g(\theta^T x_i)+(1-y_i)\log(1-g(\theta^T x_i))\right] \tag{9}$$

And for the logistic loss function of the logistic regression model, it is necessary to take the average logistic loss function over the entire training dataset, i.e:

$$\begin{aligned} J(\theta)&=\frac{1}{m}\sum_{i}^{m}Cost(g(\theta^T x_i),y_i)\\ &=-\frac{1}{m}\left[\sum_{i=1}^{m}y_i\log g(\theta^T x_i)+(1-y_i)\log(1-g(\theta^T x_i))\right] \end{aligned} \tag{10}$$

Therefore, the following relation can be obtained:

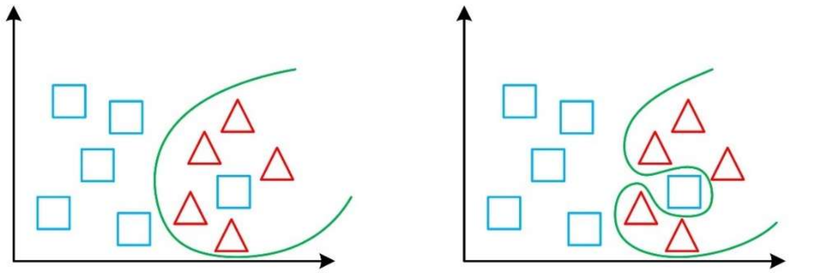
$$J(\theta)=-\frac{1}{m}l(\theta) \tag{11}$$

And our aim is to minimize the logistic loss function of the logistic regression model, even if the model's likelihood function is maximized, which is equivalent.

2.2.2. Principles of regularization. Overfitting often occurs when there is not enough training data or too many feature variables. The visualization of an overfitting model is that the complexity of the model increases as the training process proceeds, and the error on the training set decreases, but the error on the test set increases, because the trained model overfits the training set, and does not work well on the rest of the data. A model that is overfitted always fits the training set well, but this results in a model that cannot generalize to new data, so that it cannot predict new sample values very accurately.

The overfitting illustration is shown in Figure 1, the left figure shows the logistic regression model without overfitting, this logistic regression model can be generalized to new data better. And the right figure shows the logistic regression model with overfitting, you can see that this logistic regression model fits the training data very well, but it is a very distorted curve, this model will have a high error for data other than the training set and does not work well.

The overfitting problem can be solved by using regularization. Regularization does not remove any of the feature variables, but it reduces the size of the values of the parameters of the feature variables. This method is very effective, when the model has many feature variables, each of them is able to affect the prediction a little bit, so it is not desirable to remove these features, which leads to the concept of regularization.



$$g(\theta^T x) = g(\theta_0 + \theta_1 x_1 + \cdots + \theta_s x_s) \qquad g(\theta^T x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_1^2 + \theta_3 x_1^2 x_2 + \cdots)$$

Fig. 1. A graphical example of a fitting

In order to solve the overfitting problem of the model, some modifications can be made to its loss function, i.e., some terms can be added to the original loss function as shown in the following equation:

$$J(\theta) + 1000\theta_6^2 + 1000\theta_7^2 + \cdots \tag{12}$$

Or for that matter:

$$J(\theta) + 1000\theta_6 + 1000\theta_7 + \cdots \tag{13}$$

Both of these expressions are possible, and 1000 is simply a larger number specified arbitrarily. Then, if you want to minimize the loss function of the model, you must make these parameters as small as possible. Since adding these terms to the original loss function will make the new loss function very large, when minimizing this new loss function, it is important to keep the values of these parameters close to 0, as if they were ignored. Thus the desired result will be obtained, i.e., the model fits the training set well while retaining all the features, and the drawbacks associated with overfitting are avoided, and the model still works well for new data.

In the regularization method, the approach of narrowing down the values of all the parameters in the loss function of the model is taken. In fact, the smaller the values of the parameters, the simpler the corresponding model will be, the smoother the function curve will be, and thus less likely to overfitting.

2.3. Multiple logistic regression models

2.3.1. Export of models. In this paper, we consider the problem of categorizing K category, without loss of generality, assuming category $\{1, 2, \ldots, K\}$.

Assume dataset $D = \{(x_i, y_i)\}_{i=1}^n$, where $x_i = (1, x_{i1}, \ldots, x_{ip})^T \in \mathbb{R}^{p+1}$, $y_i \in \{1, \ldots, K\}, i = 1, \ldots, n$. where the first element of x_i represents the intercept term, p is the number of features, and n is the sample size.

Thus the design matrix is:

$$\mathrm{X} = \begin{pmatrix} x_1^T \\ \vdots \\ x_n^T \end{pmatrix} = (1, X_1, \ldots, X_p) \in \mathbb{R}^{n \times (p+1)} \tag{14}$$

The binary logistic regression model assumes that the logarithmic value of the ratio of the posterior probabilities of belonging to the two categories 0 and 1 is linear, i.e:

$$\log \frac{\Pr(y=1 \mid x)}{\Pr(y=0 \mid x)} = \beta^T x \tag{15}$$

The left end of the above equation is called the log odds or split logarithm, hence the logistic regression model is also called the log odds model. This model above uses category 0 as the base category.

Based on this model, generalize it to a multi-category situation. Assume that category K is used as the base class and assume that each of the $K-1$ log odds takes the form:

$$\begin{gathered}\log\frac{\Pr(y=1\mid x)}{\Pr(y=K\mid x)}=\beta_1^T x\\ \log\frac{\Pr(y=2\mid x)}{\Pr(y=K\mid x)}=\beta_2^T x\\ \vdots\\ \log\frac{\Pr(y=K-1\mid x)}{\Pr(y=K\mid x)}=\beta_{K-1}^T x\end{gathered}\tag{16}$$

where $\beta_1,\beta_2,\ldots,\beta_{K-1}\in\mathbb{R}^{p+1}$. Is thus solved:

$$\begin{gathered}\Pr(y=1\mid x)=e^{\beta_1^T x}\Pr(y=K\mid x)\\ \Pr(y=2\mid x)=e^{\beta_2^T x}\Pr(y=K\mid x)\\ \vdots\\ \Pr(y=K-1\mid x)=e^{\beta_{K-1}^T x}\Pr(y=K\mid x)\end{gathered}\tag{17}$$

And then according to:

$$\sum_{k=1}^{K}\Pr(y=k\mid x)=1\tag{18}$$

The specific form of this K posterior distribution is obtained as:

$$p_l(x,\beta):=\Pr(y=l\mid x)=\frac{\exp(\beta_l^T x)}{\sum_{j=1}^{K}\exp(\beta_j^T x)}=\begin{cases}\dfrac{\exp(\beta_i^T x)}{1+\sum_{j=1}^{K-1}\exp(\beta_j^T x)} & l=1,\ldots,K-1\\[2ex] \dfrac{1}{1+\sum_{j=1}^{K-1}\exp(\beta_j^T x)} & l=K\end{cases}\tag{19}$$

2.3.2. Parameter matrix and log-likelihood function. Define the parameter matrix as:

$$\mathrm{B}:=\begin{pmatrix}\beta^{(0)}\\ \beta^{(1)}\\ \vdots\\ \beta^{(p)}\end{pmatrix}=(\beta_1,\cdots,\beta_{K-1})\in\mathbb{R}^{(p+1)\times(K-1)}\tag{20}$$

where the row vectors:

$$\beta^{(m)}=(\beta_1^{(m)},\beta_2^{(m)},\cdots,\beta_{K-1}^{(m)})\in\mathbb{R}_{K-1},m=0,1,\ldots,p\tag{21}$$

Corresponds to the intercept term as well as p feature. And the column vector:

$$\beta_l = \begin{pmatrix} \beta_l^{(0)} \\ \beta_l^{(1)} \\ \vdots \\ \beta_l^{(p)} \end{pmatrix} \in \mathbb{R}^{p+1}, l = 1, \dots, K-1 \, (\beta_K := 0) \tag{22}$$

corresponds to each category.

Define the parameter vector as:

$$\beta := vec(\mathrm{B}) = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_{K-1} \end{pmatrix} \in \mathbb{R}^{(K-1)(p+1)} \tag{23}$$

And define:

$$\rho_m := \| \beta^{(m)} \|_1 \tag{24}$$

and:

$$\omega_m := \| \beta^{(m)} \|_2 \tag{25}$$

to measure the importance of the m st feature for classification, where $m = 1, \dots, p$.

Define the vector consisting of the first $K-1$ posterior probabilities as:

$$p(x,\beta) := (p_1(x,\beta), \dots, p_{K-1}(x,\beta))^T \in \mathbb{R}^{K-1} \tag{26}$$

It is easy to show that for any $l, k \in \{1, \dots, K-1\}$, then:

$$\frac{\partial p_l(x,\beta)}{\partial \beta_k} = \begin{cases} p_l(x,\beta)[1 - p_l(x,\beta)]x, & l = k \\ -p_l(x,\beta) p_k(x,\beta) x, & l \neq k \end{cases} \tag{27}$$

and the conditional log-likelihood function [23] with respect to the parameter vector β is:

$$\begin{aligned} \ell_n(\beta) &= \log \prod_{i=1}^{n} p_{y_i}(x_i,\beta) = \sum_{i=1}^{n} \log p_{y_i}(x_i,\beta) \\ &= \sum_{i=1}^{n} \left[\beta_{y_i}^T x_i - \log\left(1 + \sum_{j=1}^{K-1} \exp(\beta_j^T x_i)\right)\right] \end{aligned} \tag{28}$$

2.3.3. Score functions and negative Hessian matrices. The score function is the first-order derivative of the log-likelihood function with respect to the parameter vector β:

$$s_n(\beta) := \frac{\partial \ell_n(\beta)}{\partial \beta} = \begin{pmatrix} \dfrac{\partial \ell_n(\beta)}{\partial \beta_1} \\ \vdots \\ \dfrac{\partial \ell_n(\beta)}{\partial \beta_{K-1}} \end{pmatrix} \tag{29}$$

Among them:

$$\frac{\partial \ell_n(\beta)}{\partial \beta_l} = \sum_{i=1}^{n} \left[I(y_i = l)x_i - \frac{\exp(\beta_l^T x_i)}{1 + \sum_{j=1}^{K-1} \exp(\beta_j^T x_i)} x_i \right]$$
$$= \sum_{i=1}^{n} [I(y_i = l) - p_l(x_i, \beta)] x_i, l = 1, \ldots, K-1 \tag{30}$$

And so it gets:

$$s_n(\beta) = \sum_{i=1}^{n} \left(\begin{pmatrix} I(y_i = 1) \\ \vdots \\ I(y_i = K-1) \end{pmatrix} - \begin{pmatrix} p_1(x_i, \beta) \\ \vdots \\ p_{K-1}(x_i, \beta) \end{pmatrix} \right) \otimes x_i$$
$$:= \sum_{i=1}^{n} (g_i - p(x_i, \beta)) \otimes x_i \tag{31}$$

where $\otimes$ represents the Kronecker product and:

$$g_i := (I(y_i = 1), \ldots, I(y_i = K-1))^T \tag{32}$$

The Hessian matrix [24] is the second order derivative with respect to the parameter vector β. Thus the negative Hessian matrix is:

$$h_n(\beta) := -\frac{\partial^2 \ell_n(\beta)}{\partial \beta \partial \beta^T} = \left(-\frac{\partial^2 \ell_n(\beta)}{\partial \beta_l \partial \beta_k^T} \right)_{j,k=1,\ldots,K-1} \tag{33}$$

where, for any $l, k \in \{1, \ldots, K-1\}$, then:

$$\frac{\partial^2 \ell_n(\beta)}{\partial \beta_l \partial \beta_k^T} = \begin{cases} -\sum_{i=1}^{n} p_l(x_i, \beta)(1 - p_l(x_i, \beta)) x_i x_i^T, & l = k \\ \sum_{i=1}^{n} p_l(x_i, \beta) p_k(x_i, \beta) x_i x_i^T, & l \neq k. \end{cases} \tag{34}$$

So there:

$$h_n(\beta) := \sum_{i=1}^{n} V(x_i, \beta) \otimes (x_i x_i^T) \tag{35}$$

Among them:

$$V(x, \beta) = (v_{lk}(x, \beta))_{k,l \in \{1,\ldots,K-1\}} \in \square^{(K-1)\times(K-1)} \tag{36}$$

And:

$$v_{lk}(x, \beta) = \begin{cases} -p_l(x, \beta) p_k(x, \beta), & k \neq l, \\ [1 - p_l(x, \beta)], & k = l. \end{cases} \tag{37}$$

Since V is a principal diagonal dominant matrix and the diagonal elements are all non-negative, V is a semi-positive definite matrix. By Lemma 1, $h_n(\beta)$ is also a semipositive definite matrix. Thus $\ell_n(\beta)$ is a concave function of the parameter vector of β.

3. Findings and analysis of the social integration of foreign immigrants

3.1. Data analysis and processing

3.1.1. Description of raw data. The data come from the foreign immigrant application platform on the official website of a certain country, and the main factors affecting the integration of foreign immigrants are shown in Table 1. By analyzing the raw data, these factors can be divided into four perspectives, namely, “immigrant integration-related system, community integration structure, cultural identity and immigrant-native two-way perception”.

1)A Immigrant Integration Related Systems include: A1 Immigrant Integration Related Laws, A2 Disembedded Policies on Services and Guarantees for Foreign Immigrants, and A3 Lack of Assessment Mechanisms for Foreign Immigrant Integration.

2) B Community integration structure includes: B1 administrative forces are not well embedded, B2 social forces are marginalized.

3) C cultural identity: C1 home country culture in “high potential”, C2 cultural adjustment difficulties, C3 cultural differences leading to conflicts in the community neighborhood.

4) The two-way immigrant-native perceptions include: D1 “grouping” of foreign immigrants to form ethnic networks that are not embedded in the community, and D2 biased perceptions of foreign immigrants by the local population.

Table 1. The main factors influencing the integration of foreign immigrants

Primary factor	Meaning description
Immigration integration	Immigration is suspended in the law
	The exclusion of foreign immigration service protection policies
	The migration of foreign immigrants is missing
Community integration	Administrative power is embedded in exhaustion
	Social power marginalization
Cultural identity	The national culture is "high potential energy"
	Cultural adjustment is difficult
	Cultural differences lead to community conflict
"Immigration - local" two-way cognition	The "cram" foreign immigrants form a ethnic network that is embedded in the community
	Immigration is suspended in the law

3.1.2. Data visualization and analysis. Data visualization can clearly and intuitively show the relationship between independent variables and dependent variables. Through visualization analysis, it can provide data support and theoretical basis for the subsequent selection of features. Visualization analysis of the data set used this time, part of the visualization results are shown below. In this paper, each influencing factor is divided into five grades from high to low, namely “very satisfied, satisfied, average, relatively dissatisfied, extremely dissatisfied”, which are named as A, B, C, D and E. The data are then analyzed and analyzed to determine the relationship between the independent variables and the dependent variables.

1) The impact of relevant systems and community integration structures on foreign immigrant integration. The impact of the relevant system and community integration structure on the integration of foreign immigrants is shown in Figure 2, which shows that the better the relevant system, the higher the success rate of foreign immigrant integration. When analyzed in context, a foreign immigrant is more likely to immigrate if he or she has a high level of satisfaction with current immigration laws. For example, when a foreign immigrant feels very satisfied with the immigration-related laws, his or her immigration satisfaction coefficient reaches 0.44825, but when he or she is extremely dissatisfied, his

or her immigration satisfaction coefficient reaches only 0.07791, and his or her willingness to immigrate will be reduced by 82.62%. This shows that the improvement of immigration laws and regulations is crucial to the immigrants' willingness to integrate.

When the community integration structure is imbalanced, the integration rate of immigrants is significantly reduced. When analyzed in the context of practical comparisons, often the more imbalanced the community integration structure is, the lower the integration efficiency outcome of immigrants. When immigrants are very satisfied with the community integration structure, the correlation coefficient of their integration is 0.312, but when the community integration structure is imbalanced and foreign immigrants are extremely dissatisfied, the correlation coefficient of immigrants' integration into the community decreases to 0.0699.

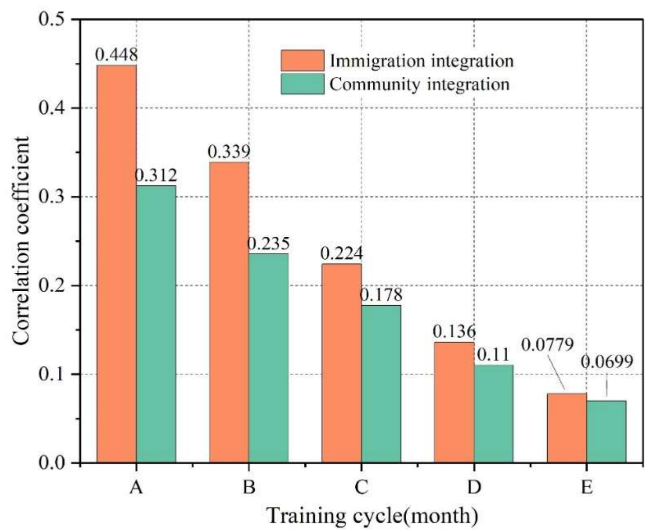


Fig. 2. The influence of system and community structure on migration

2) The effects of cultural identity and two-way cognitive bias on immigrant integration. In this paper, the degree of cultural identity and immigrant-native bidirectional cognition are divided into five grades, which are: A very much agree, B agree, C average, D disagree, E extremely opposed; cognitive bias are: A deviation is extremely large, B has a little bit of bias, C has no bias, D is more agreeable, E is very much agreeable. The effects of cultural identity and cognitive bias on immigrant integration are shown in Figure 3, which shows that when foreign immigrants have a high degree of cultural identity with the local culture, their integration is higher and faster, and accordingly, when the degree of identity is low, it is difficult for them to be integrated into new environments and collectives. For example, when the degree of cultural identification is higher (level A), the correlation coefficient of foreign immigrants' integration reaches 0.961; while when they have a high degree of cultural disidentification, the correlation probability of their immigrating to the local area and integrating is extremely small, and the correlation coefficient is only 0.0226 at this time.

The immigrant-native bidirectional perception shows that when both people have a great deviation in their perceptions of each other (level A), it is extremely difficult for both to accept each other (0.0243), but as their two sides get to know and accept each other, their perceptions of each other become deeper, and they change from initial rejection to non-acceptance but showing respect, and non-rejection as well. In the end, foreign immigrants and natives are likely to maintain a form of harmony based on mutual respect. At this point, they will go through the process of “a little bias - no bias - more agreement - a lot of agreement”, which will also increase their probability of local integration from 0.376 to 0.98.

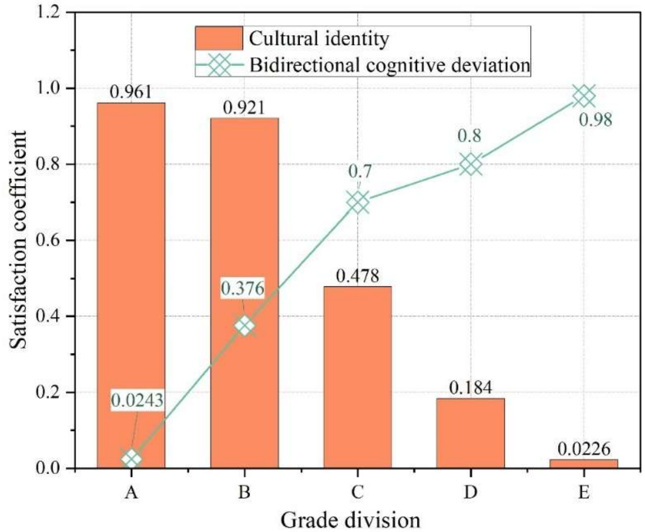


Fig. 3. Cultural identity and cognitive bias affect immigration

3.1.3. Feature correlation analysis. Feature correlation refers to the closeness of correlation between features. If there are highly correlated features, it will seriously affect the accuracy of the prediction model. Therefore, for features with high correlation, it is sufficient to choose one of the features. The correlation coefficients of variables X and Y are calculated as shown in the following equation:

$$\rho(X,Y)=\frac{E(XY)-E(X)E(Y)}{\sqrt{E(X^2)-E^2(X)}\sqrt{E(Y^2)-E^2(Y)}} \tag{38}$$

where, E denotes the mathematical expectation. The correlation coefficient matrix plot was used to analyze the feature correlation of the raw data.

The results of the correlation coefficient matrix plot are shown in Figure 4. The findings show that the highest correlation is between D1 and D2 with a correlation coefficient of 0.9214. This is followed by A2 and A1, between which there is also an extremely high correlation, with a correlation coefficient of 0.8678. The other characteristics are more or less correlated with each other, but the correlations are all below 0.7. It is not difficult to explain this in the context of the actual situation, and in general, in the actual immigration, it is also consistent with this phenomenon and pattern. For example: when the system related to immigrant integration is perfect, expatriates have a correspondingly high degree of trust in the local government, at which point they will feel at ease to understand the local culture, learn from each other and local residents, reduce prejudice against their culture, and better integrate into the post-immigration circle of life.

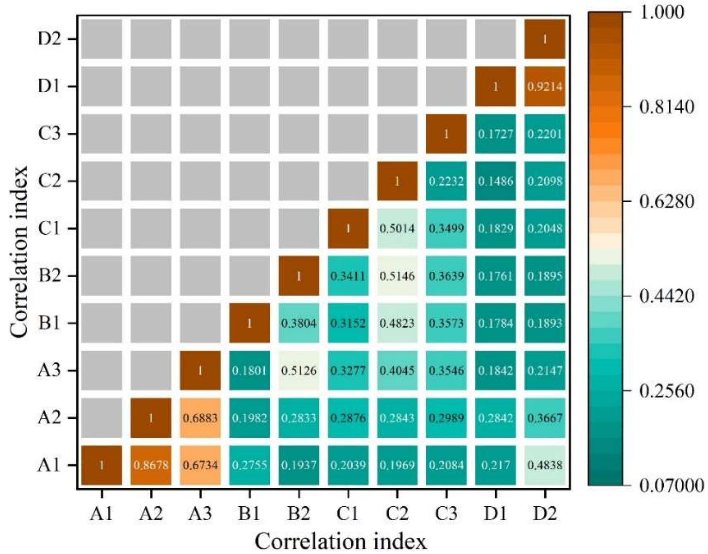


Fig. 4. Correlation coefficient matrix graph results

3.2. Multiple logistic regression

3.2.1. Multiple logistic regression modeling. The dataset chosen for this paper is composed of 5,000 expatriate migrants who migrated and settled in China from 2013 to 2023, and after data preprocessing, the resulting dataset has 10 variables. The 5000 foreign immigrants are categorized into the following four categories according to their migration purposes: work-based immigrants, study-based immigrants, investment-based immigrants, and trust-based immigrants. The dataset is divided into two parts: training set and test set. In order to find the optimal parameters, the grid search method is used for modeling, and the regularization terms from the previous section are added to output the parameters of each variable to obtain the coefficient matrix of the variables, and the regularization coefficients of the multivariate logistic regression model are shown in Table 2, where all the values are retained to 4 decimal places. From the final results obtained, the coefficients of many variables in the 10 indicators are within 1.

For the part of the population that chooses to migrate due to the nature of their work, its coefficient on D1 is the highest at 1.0071, which shows that this part of the population is very concerned about whether there is any cognitive bias in the local population towards foreign immigrants. However, they do not stick to their own ways and will take the initiative to befriend their neighbors and actively integrate into the current environment. In addition, this group is more concerned about the system related to immigrant integration and the structure of community integration, and less concerned about cultural perceptions.

For the group of foreign immigrants who are learning immigrants, they are more concerned about the cultural identity of the immigrant region (1.1568-1.8751) and the existence of cognitive bias of local residents toward foreign immigrants (1.1782), followed by the integrity of relevant laws and regulations (0.9484-1.4079), and are extremely concerned about the integration structure of the community was of minimal concern (<0.1).

For the third category of investor immigrants, their concern for the system related to immigrant integration is extremely high, with a coefficient of concern of 0.9376-2.1435, while their concern for the other three aspects, such as culture, is less than 0.1.

For the fourth type of migrants, the attention to the relevant system of immigrant integration, the structure of community integration and the two-way cognition of "immigrants and natives" is greater than 1, while the attention to cultural identity is between 0.9374 and 1.4342. Immigrants in this group are desperate for acceptance, so they tend to pay more attention to all aspects than other groups.

Table 2. The regularity of logistic regression model

Variable/ Class	Working sex immigrants	Learning immigrants	Investment immigrant	Toppless migration
A1	0.8432	0.9484	1.0672	1.8178
A2	0.9303	0.9446	2.1435	1.1802
A3	0.8487	1.4079	0.9376	1.4799
B1	0.8563	0.0798	-0.0803	1.1943
B2	0.8573	0.0972	0.0262	1.4292
C1	0.2281	1.1568	0.0392	0.9481
C2	0.5776	1.6058	0.0813	0.9374
C3	0.3108	1.8751	0.0756	1.4342
D1	-0.2687	0.5036	0.0309	3.0887
D2	1.0071	1.1782	0.0607	3.0104

3.2.2. Multiple logistic regression model evaluation. The effect of the multiclassification multiple logistic regression model was analyzed, and the confusion matrix of the multiple logistic regression is shown in Table 3, in which there are 26 in category 1, 10 correctly classified, with a poor rate of classification correctness (38.46%); 56 correctly classified in category 2, with a good rate of classification correctness (78.87%), 5 misclassified to category 1, 7 misclassified to category 3, and 3 misclassified to category 4 category. Category 3 had 59 correctly classified (71.08%) with 14 misclassified to category 2, 7 to category 4 and 3 to category 1. Category 4 had 26 correct classifications (56.52%), but 16 misclassifications to category 3 and 2 misclassifications to categories 1 and 2. It is easy to find that the classifier for category 2 is better and category 1 is worse, indicating that the multivariate logistic regression model is better for modeling immigrants who learn to migrate and not so good for modeling immigrants who work to migrate.

Table 3. The confusion matrix of logistic regression

Confusion matrix		Prediction category			
		1	2	3	4
Actual category	1	10	12	2	2
	2	5	56	7	3
	3	3	14	59	7
	4	2	2	16	26
Classification accuracy		38.46%	78.87%	71.08%	56.52%

The overall prediction accuracy of the multiclassified multiple logistic regression model was 0.768, and the model evaluation results of the multiple logistic regression on the test set are shown in Table 4. The highest precision rate is for category 4, which is 0.83, indicating that the probability of actually being 4 in the samples predicted to be category 4 reaches 0.84, which is still high in general. The highest recall and F1 are both for category 2, with values of 0.89 and 0.83, respectively, which means that learning to migrate foreign immigrants is better modeled. Overall, the results obtained from multicategorical multiple logistic regression modeling are credible.

Table 4. The model evaluation results of logistic regression

Index	1	2	3	4
Accuracy rate(%)	0.75	0.82	0.75	0.84
Recall rate(%)	0.46	0.89	0.79	0.64
F1 value	0.57	0.83	0.76	0.77

4. Conclusion

In this paper, in order to explore the barriers and facilitators of foreign immigrants' social integration, we constructed a multivariate logistic regression model optimized based on the principle of regularization to analyze the factors of foreign immigrants' social integration.

1) In this paper, the foreign group is divided into four categories, namely, immigrant integration-related system, community integration structure, cultural identity and immigrant-native two-way cognition.

2) The better the immigration-related system and the higher the degree of cultural identity of foreign immigrants, the higher the success rate of foreign immigrant integration. When there is an imbalance in the community integration structure, the integration rate of immigrants is significantly lower. When the cognitive deviation between the two sides is extremely large (level A), it is extremely difficult for both sides to accept each other (0.0243), and as the deviation decreases, the probability of their integration into the local area will be higher and there are correlations between most of the characteristics, among which the correlation between D1 and D2 is the highest (0.9214), followed by A2 and A1 (0.8678), and the correlations among the rest of the indicators are all less than 0.7.

3) This paper divides immigrants into the following four categories according to the purpose of their migration: the first category of work-based immigrants in this group is more concerned about immigrant integration-related systems and community integration structures, and less concerned about cultural cognition. The second category of learning migrants is mainly concerned with cultural identity and the existence of cognitive bias of local residents towards foreign migrants. The third category of investor immigrants is more concerned with laws and regulations. The fourth category of sheltered immigrants is more concerned about all aspects of laws and regulations and culture. The accuracy of the multivariate logistic regression model for the division of the four groups was 38.46%, 78.87%, 71.08% and 56.52%, respectively, and the results obtained by multiclassified multivariate logistic regression modeling have credibility.

Funding

This work was supported by the Social Science Research Program of the Department of Education of Jilin Province, China: ‘Research on the Management of Social Integration of Foreigners in Jilin Province’ (Project No. JJKH20231066SK); the Educational Teaching Reform Program of Jilin Police College: ‘Research on the Innovation of Internationalized Talent Cultivation Mode in Local Public Security Colleges and Universities’ (Project No. JJY-018).

References

[1] Laurentsyeva, N., & Venturini, A. (2017). The social integration of immigrants and the role of policy–A literature review. *Intereconomics*, 52(5), 285-292.

[2] Gidron, N., & Hall, P. A. (2020). Populism as a problem of social integration. *Comparative Political Studies*, 53(7), 1027-1059.

[3] Hainmueller, J., Hangartner, D., & Pietrantuono, G. (2017). Catalyst or crown: Does naturalization promote the long-term social integration of immigrants?. *American Political Science Review*, 111(2), 256-276.

[4] McMillan, C. (2019). Tied together: adolescent friendship networks, immigrant status, and health outcomes. *Demography*, 56(3), 1075-1103.

[5] Raihan, M. M., Chowdhury, N., & Turin, T. C. (2023). Low job market integration of skilled immigrants in Canada: the implication for social integration and mental well-being. *Societies*, 13(3), 75.

[6] Alegría, M., Álvarez, K., & DiMarzio, K. (2017). Immigration and mental health. *Current epidemiology reports*, 4, 145-155.

[7] Kancs, D. A., & Lecca, P. (2018). Long‐term social, economic and fiscal effects of immigration into the EU: The role of the integration policy. *The World Economy*, 41(10), 2599-2630.

[8] Davidov, E., & Semyonov, M. (2017). Attitudes toward immigrants in European societies. *International Journal of Comparative Sociology*, 58(5), 359-366.

[9] Plenty, S., & Jonsson, J. O. (2017). Social exclusion among peers: The role of immigrant status and classroom immigrant density. *Journal of youth and adolescence*, 46, 1275-1288.

[10] Fouka, V., Mazumder, S., & Tabellini, M. (2022). From immigrants to Americans: Race and assimilation during the Great Migration. *The Review of Economic Studies*, 89(2), 811-842.

[11] Kaushik, V., & Drolet, J. (2018). Settlement and integration needs of skilled immigrants in Canada. *Social Sciences*, 7(5), 76.

[12] Simpson, J. (2021). ESOL, integration and immigration. *NYS Tesol Journal*, 5-17.

[13] Gudrun Jensen, T., Weibel, K., & Vitus, K. (2017). 'There is no racism here': Public discourses on racism, immigrants and integration in Denmark. *Patterns of Prejudice*, 51(1), 51-68.

[14] Gretter, A., Machold, I., Membretti, A., & Dax, T. (2017). Pathways of immigration in the Alps and Carpathians: Social innovation and the creation of a welcoming culture. *Mountain Research and Development*, 37(4), 396-405.

[15] Johnston, J., & Audunson, R. (2019). Supporting immigrants' political integration through discussion and debate in public libraries. *Journal of librarianship and information science*, 51(1), 228-242.

[16] Steinhardt, M. F. (2018). The impact of xenophobic violence on the integration of immigrants (No. 11781). *IZA Discussion Papers*.

[17] Vahan, S. (2018). Social integration of immigrants and the attitude of the native population in European countries. *CERGE-EI Working Paper Series*, (629).

[18] Mei, L., Liu, K., Xiong, L., & Zhu, B. W. (2022). Approach Based on Data Mining for Exploring the Hierarchical Decision-Making Rules between the Generation of Transnational Immigrants' Sense of Place in Religious Spaces and Perception of the Environment. *Religions*, 13(3), 195.

[19] Bhuyan, R., Jeyapal, D., Ku, J., Sakamoto, I., & Chou, E. (2017). Branding 'Canadian experience'in immigration policy: Nation building in a neoliberal era. *Journal of International Migration and Integration*, 18, 47-62.

[20] S. Pradeepa,Elizabeth Jomy,S. Vimal,Md. Mehedi Hassan,Gaurav Dhiman,Asif Karim & Dongwann Kang. (2024). HGATTLR: transforming review text classification with hypergraphs attention layer and logistic regression. Scientific Reports(1),19614-19614.

[21] Yunran Zhao & Xuelei Li. (2024). Better approximation of sigmoid function for privacy-preserving neural networks. Journal of Physics: Conference Series(1),012007-012007.

[22] Guo Jiajie. (2024). Research on Mobile E-commerce Recommendation Algorithms Based on Logistic Regression Improved Model Features. Academic Journal of Engineering and Technology Science(5),

[23] Sadok Ibrahim,Masmoudi Afif & Zribi Mourad. (2023). Integrating the EM algorithm with particle filter for image restoration with exponential dispersion noise. Communications in Statistics - Theory and Methods(2),446-462.

[24] Jinlong Li,Ling Wang,Yingying Wan,Kai Yang & Lin Luo. (2024). An adaptive weighted width extraction method based on the Hessian matrix for high-precision detection of laser stripe centers in low-exposure. Optics and Lasers in Engineering108436-108436.