Research on Floating Population and Enterprise Innovation Based on LightGBM

Ying Chen[1, *], ShunJie Xu[2]

[1]Nanchang Hangkong University, School of Economics and Management Nanchang, China

[2]Beijing Jiaotong University, School of Economics and Management, Beijing, China

*Corresponding author: chenying220216@163.com

Abstract. The research on the relationship between the floating population and social-economic development has always been a key research issue in population economics. This study first introduced the LightGBM to examine multi-dimensional floating population characteristics can predict enterprise innovation. Taking the dynamic survey data of China's floating population from 2011 to 2018 and listed enterprises as samples, the study verifies whether the characteristics of a floating population can predict enterprise innovation. Further, it analyzes the characteristics of the floating population and its prediction model, which has an excellent ability to predict enterprise innovation. The researchers have found that: 1) On the whole, the prediction ability to float population characteristics to enterprise innovation is limited. 2) Among many characteristics, the prediction ability of population mobility dimensional characteristics is higher. 3) There is a nonlinear relationship between characteristics and enterprise innovation, and there is a population crowding effect, which is consistent with the previous theory. 4) It is tested that the population crowding effect is through mechanisms such as human capital. This study uses the machine learning method to explore the characteristics of floating population more comprehensively and inspires the formulation and implementation of the government's industrial and talent policies and the development decisions of enterprises.

Keywords: machine learning; lightgbm; floating population; enterprise innovation.

1. Introduction

The relevant research focuses on the macro level, mainly to explore the impact of a single index change caused by the floating population on technological innovation in the region, such as rising labor costs, advanced human capital structure, the increase in the number of floating populations, etc [1-3]. In comparison, there are few studies on the impact of a more comprehensive analysis of the characteristics of the floating population at the micro-level, such as the impact of corporate technological innovation. Meanwhile, the research on floating population and enterprise indicators only pays attention to the causal relationship between them, ignoring the ability to float population characteristics to predict enterprise indicators. However, the characteristics of the floating population that can explain the enterprise indicators may not accurately predict the enterprise indicators, so it is impossible to conclude the prediction effect from the causal inference estimation results of the traditional measurement. So, can the floating population's characteristics help predict the innovation ability of enterprises? Which characteristics are more important to predict the innovation ability of enterprises? What is its prediction mechanism? There are no clear conclusions on these problems in previous studies on the floating population. Comparing the prediction ability of predictive models with different variables can reflect the ability to theoretically interpret phenomena, reveal complex laws in the data, and help to test previous conclusions and put forward to new theories [4]. This paper aims to explore the prediction effect of floating population characteristics by selecting the index of enterprise innovation to contribute to this field.

From the perspective of resource factors, the mobility of production factors will have a specific influence on production efficiency. Some scholars have concluded that the direction of human capital flow is entirely consistent with population mobility [5]. The dimensions of these floating population characteristics can be classified into the following five categories. First, in terms of population mobility, Cui and Chen [3] used the scale of population mobility as a substitute variable for population mobility.

They got that population mobility can improve the overall level of human capital and positively affect urban innovation. The other is labor cost. Dai found that the rise of labor costs significantly affects enterprise innovation [6]. The third is the structure of human capital, which is mainly reflected in education. Hu concluded that upgrading human capital structure plays various roles in promoting regional innovation ability [2]. The fourth, the age of the labor force, under the trend of population aging, the average age of the floating population dominated by the young labor force must increase, and population mobility has a significant effect on the difference of regional population aging [7]. The fifth is the residence time of the floating population, which may affect enterprise innovation. Gu [8] found that the longer the residence time of the floating population, the stronger the willingness to stay in the inflow area for a long time, which affects the local economic development. It can be seen that the previous studies on floating populations mainly focus on specific types of characteristics but lack a comprehensive comparison of the relative importance of floating population characteristics. In addition, population mobility is a concept and phenomenon produced under China's national conditions, which makes it difficult for China's floating population research to learn from other research conclusions directly. Furthermore, former studies have found a nonlinear relationship between population mobility and regional and regional enterprise innovation [6, 9]. At the same time, the simple linear fitting model is weak to explain the complex relationship between variables. So, it is necessary to consider the impact of floating population characteristics on enterprise innovation more comprehensively.

The main contents of this paper are as follows. Firstly, the machine learning method is applied to study the problem of floating population and enterprise innovation in China, and to evaluate the prediction ability to float population characteristics to enterprise innovation as a whole as well as expand the research perspective of population characteristics and innovation ability, filling and enriching the theory in the field of "Floating population relationship with innovation ability." Secondly, the LightGBM regression algorithm is used to avoid the defects of the traditional linear model and better analyze the nonlinearity and the interaction between variables. Thirdly, explore the importance of different floating population characteristics for predicting enterprise innovation capability, and analyze the relatively important floating population characteristics for the prediction mechanism of enterprise innovation in diverse regions.

2. Data source and variable description

2.1 Data and Variable Description

In this paper, the period of the data is from 2011 to 2018. For the enterprise sample, Chinese A-share listed companies are selected. In terms of urban samples, the macro panel data of 28 provinces are picked. The data processing is shown in Figure 1. After processing, this study obtains 5128 observations.

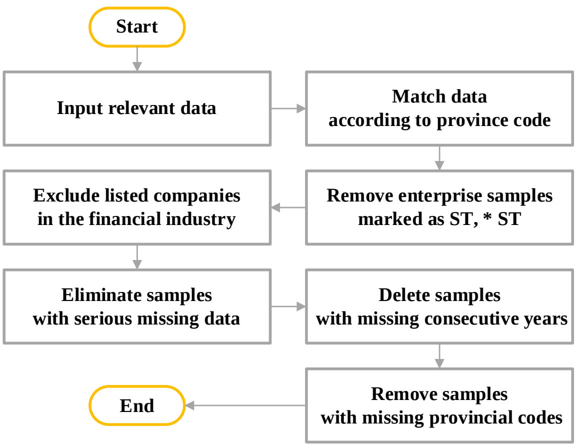


Figure 1. Data processing flow

Tibet Autonomous region, Ningxia Hui Autonomous region and Hainan Province are not included in the scope of the study because of the small sample size and the lack of too much enterprise data. The number of patent applications is used to measure the innovation ability of innovative subjects. First of all, technological innovation is the ultimate embodiment of resource input and convenience efficiency. Therefore, the number of patent applications representing innovation output can better reflect the innovation ability of the innovation subject [10]. Secondly, the patent grant requires testing and annual fees. The uncertainty, instability, and vulnerability are studied to bureaucratic factors [11]. Therefore, the number of patent applications can more truly reflect the level of innovation than the amount granted. Moreover, the core explanatory variable is the multi-dimensional characteristics of the floating population summarized in the introduction. In addition, due to the small sample size in the northeast region, in order to facilitate the research, Shandong and Liaoning are listed as the eastern region, Heilongjiang and Jilin are classified as the central region with reference to the regional classification method of the Chinese Science and Technology Statistical Yearbook. The meaning and measure of each variable are listed in Table I.

Table 1. The Meanings, Measurements, And References of Variables

Variable	***Meaning and measurement***	***Data source***	***Reference***
LnPatent	The innovation ability of the enterprise. The natural logarithm of the number of patent applications.	China Stock Market & Accounting Research Database (CSMAR),	[12-13]
PFS	The scale of population mobility. The proportion of the difference between the resident population and the registered population to the resident population.	China population and Employment Statistical Yearbook	[14]
LnCPCM	The scale of inter-provincial floating population. The natural logarithm of inter-provincial mobility in the floating population.	China Migrants Dynamic Survey (CMDS)	[15]
LnWage	Labor costs. The natural logarithm of the average monthly wage of the floating population.	CMDS	[6]
LnHCS	Human capital structure. The natural logarithm of a college degree or above in the floating population of each province.	CMDS	[2]
LA	The age of the labor force. The age of the floating population as of the time of the survey.	CMDS	[16]
RT	Length of stay. Take the floating time of the floating population minus the survey time.	CMDS	[8]
LnES	The size of the enterprise. The natural logarithm of total assets.	CSMAR	[6]
PPE	Fixed assets ratio. The ratio of net fixed assets to total assets.	CSMAR	[13]
EN	Enterprise ownership. It is classified according to the equity nature of the CSMAR database.	CSMAR	[1, 3]
LnUR	Urbanization. The natural logarithm of the proportion of the urban population to the total population.	China Statistical Yearbook, China Urban Statistical Yearbook	[17]
LnUEL	The level of the urban economy. The natural logarithm of regional gross domestic product.	China Statistical Yearbook, China Urban Statistical Yearbook	[13]
LnCS	The size of the city. The natural logarithm of the population of each province at the end of the year.	China Statistical Yearbook, China Urban Statistical Yearbook	[3]

From the causal inference, the above variables are faced with the challenge of endogenous problems. The way of dealing with the endogenesis of variables is also one of the important differences between this paper and most of the previous studies on floating population and enterprise innovation. Unlike the explanatory modeling that focuses on causality, the predictive modeling in this paper mainly explores the correlation rather than causality, so the evaluation of the prediction effect will not be affected by the endogenous existence of variables [4, 18].

2.2 Descriptive Statistics

Table II lists the descriptive statistics of predicted variables, floating population characteristics, enterprise-level, and urban-level variables. For some missing values, the neighborhood mean is used to fill it. In order to prevent the statistical deviation caused by the extreme value, this paper carries out 1% and 99% Winsorize processing on some continuous variables without affecting the distribution of the original data as far as possible. The preceding "Ln" means that the variable is processed with a natural logarithm.

According to the descriptive statistical results and changing trends in the table, combined with the detailed data of the sample, it can be found that the standard deviation of patent variable (LnPatent) of enterprises is up to 1.346, indicating that there are some differences in innovation capability among enterprises. The median population mobility scale (PFS) is 0.095, signifying that the resident population in most cities is larger than the registered population, and the floating population is an imperative force in urban construction. The maximum proportion of the inter-provincial floating population (LnCPCM) is 4.605, which is mainly reflected in the province-level municipality in the eastern regions with the largest foreign population, such as Beijing, Shanghai, and Tianjin, while the minimum value is 2.161, mainly in the provinces of Anhui, Shandong, and Hunan. From the median, maximum, and minimum labor cost (LnWage), it can see that the average monthly wage of most floating populations is on the high side, which means the increase in labor cost. The average and median of human capital structure (LnHCS) indicate that the education level of the inflow population is different. The median age of the labor force (LA) is 34, indicating that most of the floating population are in the age group with innovative ability, but the overall age structure is aging. The standard deviation of residence time (RT) is 1.100, and there are differences among floating populations in various periods and regions.

Table 2. Descriptive Statistical Results of Sample Variables

Variable	***Minimum value***	***Maximum value***	***Average value***	***Standard deviation***	***Median***
LnPatent	0.693	9.473	3.259	1.346	3.091
PFS	-0.186	0.874	0.112	0.177	0.095
LnCPCM	2.161	4.605	3.858	0.718	4.201
LnWage	7.463	8.942	8.237	0.272	8.248
LnHCS	1.348	3.641	2.513	0.504	2.492
LA	30.229	41.620	33.943	1.705	33.770
RT	2.698	8.570	4.733	1.100	4.544
LnES	17.019	28.520	22.340	1.296	22.131
PPE	0.001	0.876	0.225	0.142	0.196
EN	1.000	7.000	2.980	1.589	4.000
LnUR	3.554	4.495	4.134	0.200	4.161
LnUEL	16.631	28.389	19.793	1.045	19.749
LnCS	6.342	9.337	8.672	0.563	8.732

3. Research Methods and Model Construction

3.1 GBDT Regression Tree

The machine learning method used in this paper is the LightGBM algorithm, which is essentially an improved algorithm of Gradient Boosting Decision Tree (GBDT). Multiple decision trees integrate it, and the accumulation of the results of all decision trees is taken as the final answer, which combines the decision tree and integration idea in machine learning, effectively avoids the over-fitting problem of a single decision tree, and has high fitting accuracy in and out of samples [19].

Generally speaking, the core idea of GBDT is to reduce the loss function by constantly learning the residuals of the previous tree. A loss function mainly describes the accuracy of GBDT, and the loss function selected in this paper is a square error because it is most commonly used in the analysis of continuous result variables, and it is easier to derive mathematically. The core algorithm steps adopted in this paper are as follows [20]. It is supposed $\{x_i, y_i\}_{i=1}^{N}$ to be a training sample, and the set of predictive independent variables is x. $F(x,p)$ is an approximate function of the dependent variable y. The GBDT iterative method comprises M different and independent decision trees $h(x,a_1),...,h(x,a_m)$, and $P=\{\beta_i,\alpha_i\}_{i=1}^{M}$ is a representative parameter of this set. β is the weight of each classifier and α is a parameter of the classifier. Ba new prediction function is obtained after iteration y taking the loss function in the direction of $F_{m-1}(x)$ gradient as the input value of the following regression function.

$$F_m(x) = F_{m-1}(x) + \sum_{m=1}^{M} \beta_m h\left(x, \alpha_m\right) \tag{1}$$

From the above formula, it can be concluded that the model is constructed by using a segmented directional gradient to improve the decision tree and is constantly iterated and updated by minimization. Meantime, the LightGBM algorithm with improved GBDT usually has better performance than the commonly used machine learning methods.

3.2 Model Building

This paper makes a comprehensive consideration of the characteristics of China's floating population. First of all, a benchmark model is constructed, which includes only the variables at the enterprise level and the urban level, but does not include the characteristics of the floating population. The benchmark model consists of six control variables that describe the characteristics of the enterprise and the city. The variables at the enterprise level include enterprise size (LnES), fixed assets ratio (PPE), and enterprise ownership (EN). The variables at the urban level include urbanization (LnUR), urban economic level (LnUEL), and urban size (LnCS). Based on the benchmark model, the characteristic family of the floating population is added to construct a prediction model including the characteristics of the floating population, which is referred to as "benchmark model" and "floating population model," respectively. The two models are shown in Table 3.

Table 3. Model Definition

Predicted Variable	***Model***	***Prediction characteristic variable***
Enterprise innovation	Benchmark model	LnES, PPE, EN, LnUR, LnUEL, LnCS
	Floating population model	PFS, LnCPCM, LnWage, LnHCS, LA, RT, LnES, PPE, EN, LnUR, LnUEL, LnCS

The models take the enterprises as the research object, and the annual data of each enterprise will be taken as a sample in the model construction. Then, the annual characteristics of each enterprise can be used as a set of input variables of the model. It is drawn lessons from the practice of Fan [21]. Suppose the enterprise vector: $\overrightarrow{E_j^t}=\left(x_1^t, x_2^t, \ldots, x_n^t\right)$. x_i^t is the i characteristic that affects the innovation ability of enterprises in the t-year, $t=2011,2012,\ldots,2018$; $i=1,2,3,\ldots,n$; n is the number of features that affect the innovation ability of enterprises; y_j^t is the innovation capability of the j enterprise in the t-year, $t=2011,2012,\ldots,2018$; $j=1,2,3,\ldots,n$; The input variable is: $\overrightarrow{Y_j^t}=y_j^t$.

After calibrating the model with the samples of known enterprise innovation capability, the enterprise innovation capability of unknown samples can be predicted by the above formula. Based on the above theory, the model's main steps are shown in Figure 2.

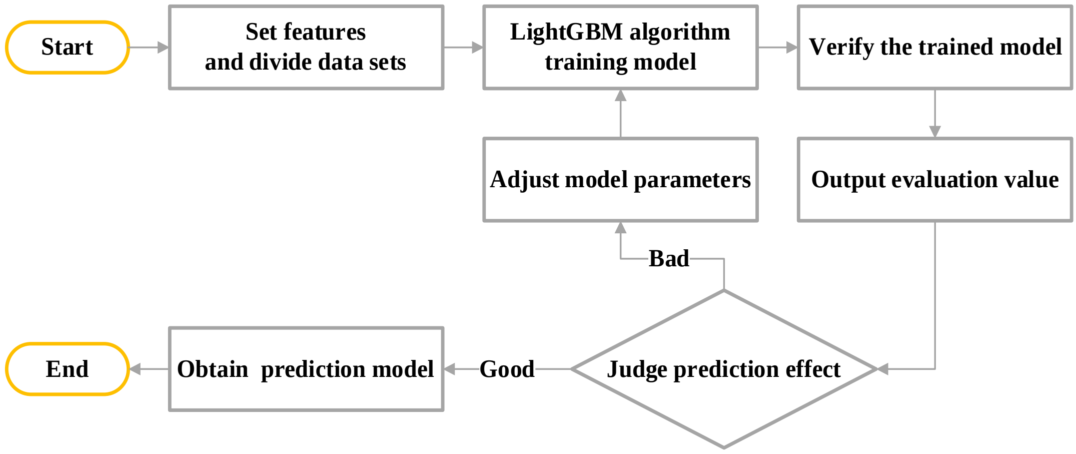


Figure 2. Model building

In the process of building the model, it is worth noting that: many previous studies are divided directly according to proportion, but only one group of the training set and test set are selected, which may lead to a great difference in the prediction performance of the model between the training set and the test set, and the prediction accuracy deviates from the actual value [22]. Therefore, this paper relies on experience to introduce k the (10) fold cross-validation method [23] to improve information utilization of samples and make the prediction results stable and accurate.

4. Empirical test and result analysis

4.1 Can the Floating Population Characteristics Predict the Innovation Ability of Enterprises?

To empirically test whether the characteristics of the floating population can greatly improve the prediction effect of the model on the innovation ability of enterprises, this part uses the trained forecasting model and the samples of Chinese A-share listed companies. Table IV lists the fitting effect of the "LightGBM-floating population model" and adds the "LightGBM-benchmark model," which includes variables at the enterprise level and urban level as a comparison. By analyzing the degree of improvement of the fitting effect of the floating population model compared with the benchmark model, it can judge the contribution of the characteristics of the floating population as a whole to the prediction of enterprise innovation capability. Meanwhile, the "OLS-benchmark model" and "OLS-floating population model" are taken as the control group to test whether the machine learning method of LightGBM can significantly improve the model's prediction ability.

The evaluation index is selected regarding the research method of Lu [18]. The goodness of fit R_{s1}^2 in the sample is calculated from the fitting results of the training set. The external goodness-of-fit R_{s2}^2 and external mean square error MSE_{s2} are calculated from the prediction results of the test set.

In the beginning, the fitting effect of LightGBM in Table 4 is much better than that of the OLS model, which shows that LightGBM is better than OLS in both training set fitting ability and out-of-sample fitting ability. For example, in the comparison of floating population models, the index of R_{s1}^2 and R_{s2}^2 of LightGBM is 34.758% and 24.019%, which is higher than that of OLS, respectively. It is proved that Liang [24] and other scholars proposed that traditional linear models are poor in handling nonlinear relationships or high-dimensional regression problems in complex systems. The tree-based nonlinear model of LightGBM can effectively carry out feature cross-effects and improve fitting ability.

Table 4. Model Fitting Result

Evaluation Index	*Prediction Model*	*Benchmark Model*	*Floating Population Model*	*Improve Effect*
R_{s1}^2 (%)	LightGBM	61.566	64.597	3.031
	OLS	27.713	29.839	2.126
R_{s2}^2 (%)	LightGBM	57.668	60.155	2.487
	OLS	34.875	36.136	1.261
MSE_{s2}	LightGBM	0.904	0.851	-0.053
	OLS	1.422	1.394	-0.028

In addition, compared with the characteristics at the enterprise level and the urban level, the floating population characteristics play a limited role in predicting the innovation capability of China's A-share listed companies. Combining the results of various indicators, it is found that the enormous improvement effect is that the R_{s1}^2 is floating population characteristics on enterprise innovation capability is 3.031%, which is much lower than that of enterprise-level and city-level characteristics on enterprise innovation capability prediction (61.566%). It can be seen that although the floating population can improve the model to predict the performance of enterprise innovation, the ability to predict enterprise innovation is limited as a whole. The possible reason is that, first of foremost, the proportion of the floating population with economic activities as the main floating purpose has consistently accounted for more than half [25], of which migrant workers account for a relatively large proportion [26]. They generally have a low level of education and skills, and their jobs are concentrated in labor-intensive low-end production links, which have low requirements for technological innovation, strong substitutability, and high job mobility [27]. Secondly, the emergence of new economic forms, such as platform economy, also impacts the outflow of migrant workers out of the manufacturing industry. More and more migrant workers are turning from original factory workers to networkers [28]. Some scholars have advised guarding against the threat of this de-skilled gig economy to the manufacturing industry [29]. Finally, the impact of the floating population on enterprise innovation is a long-term accumulation process, which should be realized through the interaction of the association mechanism and the characteristics of the floating population at the enterprise level and urban level. It will be analyzed and introduced below.

4.2 Analysis on the Importance of Floating Population Characteristics

This paper uses the SHAP model [30] to quantify the impact of floating population characteristics on enterprise innovation capability and the interaction between characteristics. Figure 3 is a summary diagram of SHAP, which visually shows the importance of variables in the floating population model for predicting enterprise innovation. The SHAP model produces a predicted value for each prediction sample, and the SHAP value is assigned to each feature in the sample. From the LnES to the LA, the degree of importance decreases successively. The SHAP value shows the distribution of the influence of each feature on the output of the model, in which dark color represents high eigenvalue, and light color means low eigenvalue.

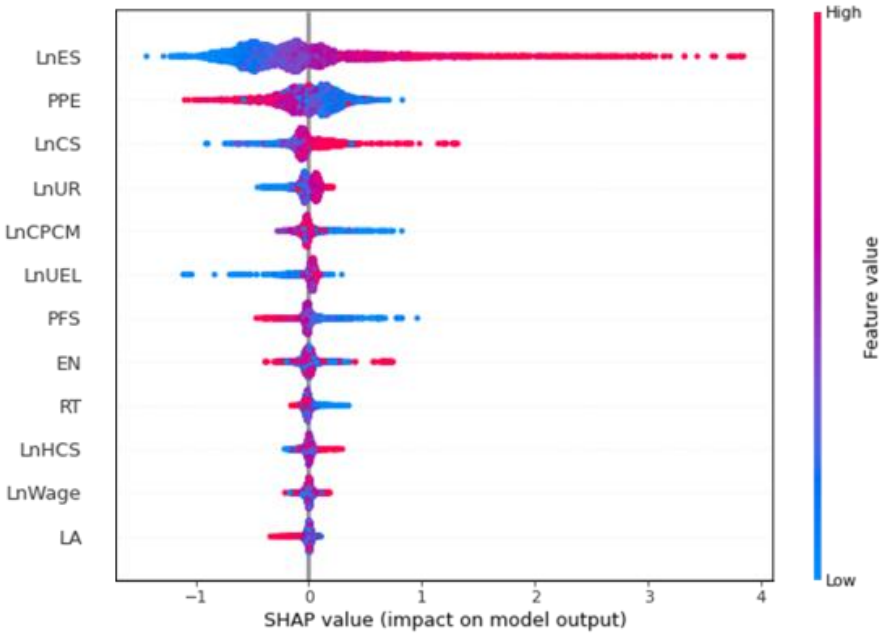


Figure 3. SHAP characteristic analysis

It can be found that among the characteristics of floating population, LnCPCM and PFS have a significant impact on the model, followed by RT, LnWage, LnHCS, LA. In addition to EN, LnES and PPE have the greatest impact on the model. In the characteristics of the urban level, the effects of LnCS, LnUR, and LnUEL are significant. Based on the order of feature importance, except for individual outliers, LnES, LnCS, LnUR, LnUEL, and LnHCS positively impact enterprise innovation. Specifically, the higher the value of these characteristics, the greater the SHAP value, and the higher the enterprise innovation ability corresponding to it, which also verifies the theory of relevant scholars [2, 17, 30–33]. Among the characteristics of the floating population, only LnHCS has a positive impact on enterprises. For improving the innovation ability of different regions and enterprises, upgrading human capital structure is more and more important [2, 32]. All urban characteristics play a positive role, indicating that the higher the overall level of urban development, the more sound the collaborative innovation network is and conducive to promoting enterprise innovation [34].

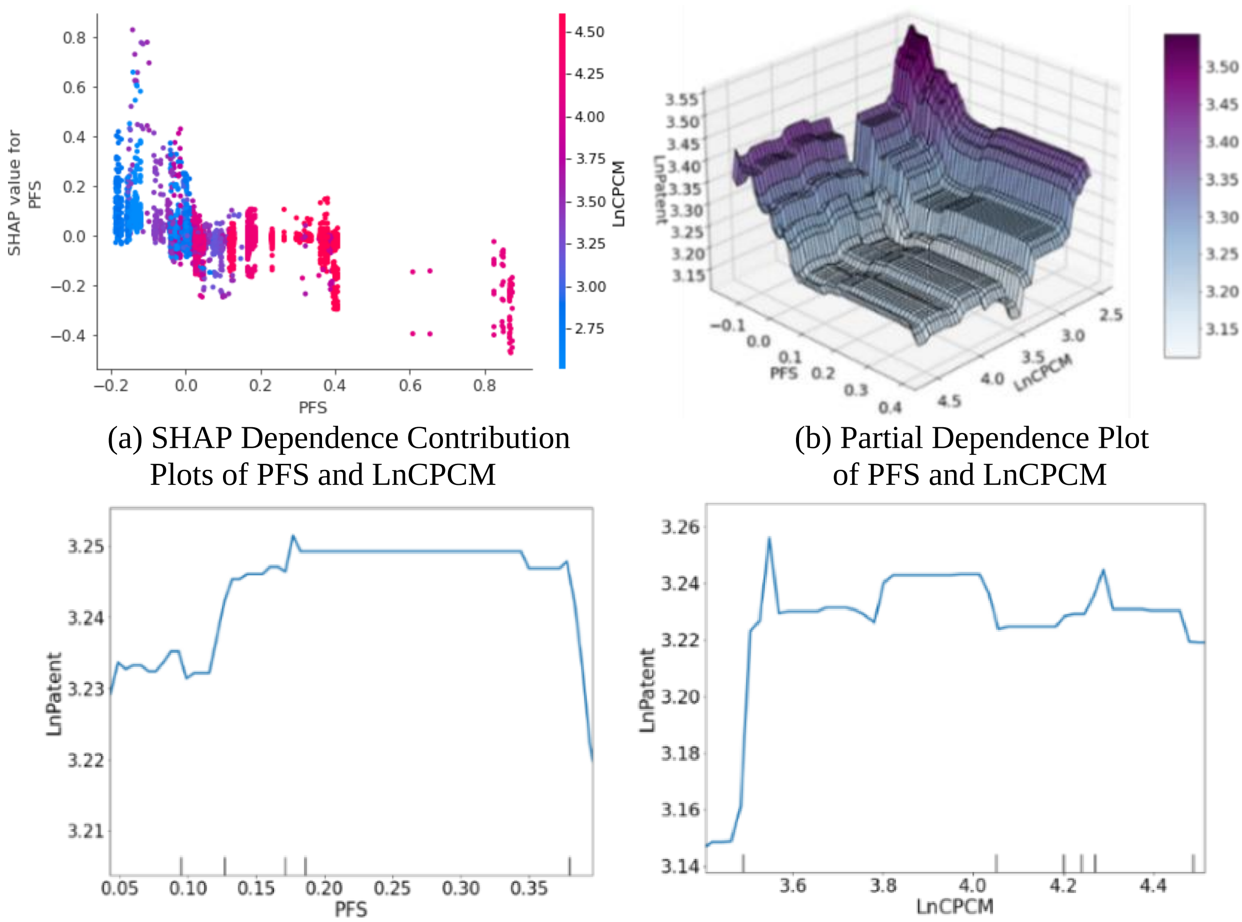
On the contrary, PPE, LnCPCM, PFS, RT, and LA have significant negative power on the innovation capability of enterprises. With the increase of eigenvalues, the SHAP value decreases, making enterprises' innovation capability lower. The characteristics of LnCPCM and PFS show that when the inflow of floating population makes the size of the urban population beyond the reasonable range, the promoting effect of population mobility on the innovation ability of inflow will be weakened and even produce crowding effect [9]. The RT and LA characteristics changes show that most floating populations live briefly and are mainly young and middle-aged. With the aging, although they have obtained the qualification of long-term residence, they are in the age stage of low labor ability and innovation ability. The impact of EN and LnWage on enterprise innovation is both positive and negative, and there are great differences in innovation capability among non-state-owned enterprises in enterprise ownership. In contrast, the positive impact of rising labor costs on enterprise innovation is more than the negative, which is related to the internal factors of the enterprise and the region [1, 6].

4.3 Prediction Pattern of Important Floating Population Characteristics for Enterprise Innovation

After analyzing the importance of variables, this study found that the scale of population mobility and the proportion of the inter-provincial floating population are the two most important factors to predict the innovation ability of enterprises in the characteristics of the floating population. In order to explore the relationship between these two characteristics and enterprise innovation, this part first exploits the SHAP dependence contribution plots to examine the prediction model of the proportion of the inter-provincial floating population and the scale of population mobility for enterprise innovation and further introduces the Partial Dependence Plot (PDP) as a robustness test [35].

As shown in Figure 4 (a), when the eigenvalue of PFS changes from 0.0 to 0.8, the value of LnCPCM increases and the SHAP value of PFS decreases. Under the interaction of the two characteristics, the larger the scale of population mobility is, the higher the proportion of the inter-provincial floating population will lower enterprises' innovation ability. Ignoring the color of the graph, but focusing on the influence of the changing PFS on the model output, it is found that with the increase of the PFS value, the SHAP value remains at a low level after decreasing, which agrees that there is a population crowding effect proposed by scholars [9, 17, 35, 36]. As depicted in Figure 4 (b), The influence trend of PFS and LnCPCM characteristics on enterprise innovation capability is consistent with the trend of SHAP dependence contribution plots. Furthermore, it is shown the interval effect, which shows that the SHAP value method is robust to the model recognition results.

Meantime, the positive impact of the change of PFS single feature changes from 0.05 to 0.2 in Figure 4 (c) on enterprise innovation increases, and it demonstrates the obvious inhibitory effect on enterprise innovation only after the value up to 0.35, that is, when the scale of population mobility reaches 35% it will have an apparent inhibitory effect on enterprise innovation. LnCPCM single feature is positively correlated with enterprise innovation from 3.4 to 3.6. Therefore, it shows an obvious negative correlation after 4.5. It proves an interaction between the LnCPCM feature and the PFS feature.



(a) SHAP Dependence Contribution Plots of PFS and LnCPCM

(b) Partial Dependence Plot of PFS and LnCPCM

(c) Partial Dependence Plot of PFS

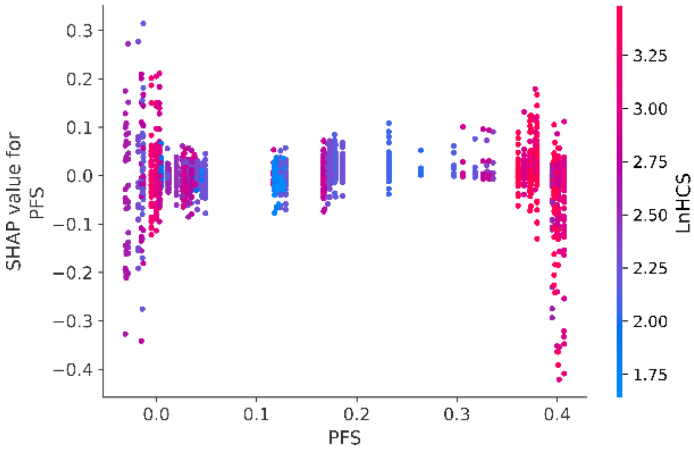
(d) Partial Dependence Plot of LnCPCM

Figure 4. SHAP Dependence Contribution Plots and Robustness Test Based on PDP

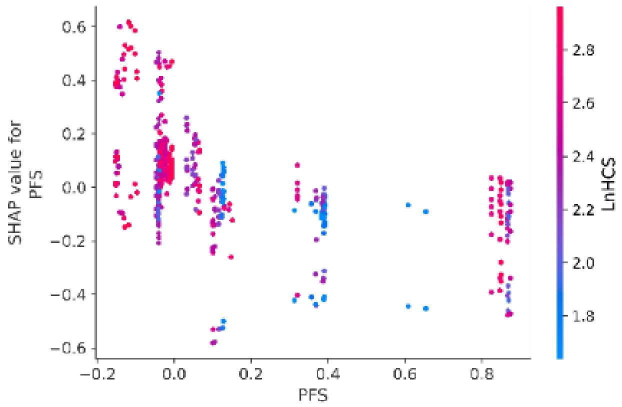
For the population crowding effect in the above analysis, the existing highly relevant research explanations are mainly verified by the human capital mechanism. Moreover, a sub-regional analysis is carried out because related studies show great differences in human capital in different regions.

1) Verify the influence of the human capital mechanism.

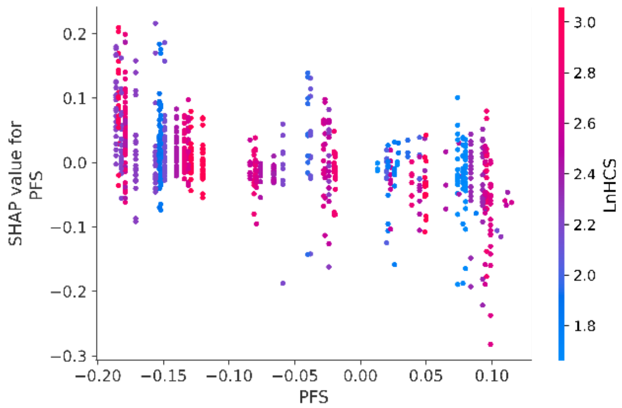
In Figure 5, the population crowding effect is reflected in varying degrees in all three regions. Figure 5 (a) depicts that when PFS ranges from 0 to 0.1 in the eastern region, the higher the eigenvalue of LnHCS, the greater the SHAP value of the PFS feature. In other words, when the scale of population mobility is low, upgrading the human capital structure of the floating population has a significant positive impact on the innovation ability of enterprises. When the PFS is from 0.3 to 0.4, the LnHCS eigenvalue increases, while the SHAP value of the PFS feature decreases. It is shown that when the scale of the floating population continues to expand, the upgrading of human capital structure does not contribute to enterprise innovation.



(a) Eastern Region



(b) Western Region



(c) Central Region

Figure 5. Regional SHAP Dependence Contribution Plots of PFS and LnHCS

Furthermore, through the sample, the eigenvalues of PFS in the range of 0.3 to 0.4 are mainly in the areas where talents are saturated, such as Beijing, Shanghai, and Tianjin. Some of the floating population in these areas are unemployed or engaged in low-skilled labor, resulting in the phenomenon of "false high" of human capital [36]. In figure 5 (b), when the eigenvalues of PFS in the western region are in the range of -0.2 to 0, the higher eigenvalues of LnHCS make the SHAP values of PFS features large. Through further observation of the samples, it is found that the specific regions of PFS eigenvalues in this interval are mainly Chongqing, Sichuan, and Shaanxi, which are not only the labor export regions but also the regions with the most substantial technological innovation capability in the western region [37]. Therefore, the inflow of a high-quality population has a more significant positive impact on the technological innovation of enterprises in these areas, while the Xinjiang region with a PFS of 0.8 is affected by its single industrial structure. Human capital does not play an innovative role in the region.

When the PFS value in the central region of figure 5 (c) varies from -0.2 to -0.05, the level of LnHCS eigenvalues has little effect on the SHAP value of PFS features but changes after 0.05 and shows a similar crowding effect. From the sample, it can be found that the population mobility scale less than 0 is the major provinces with population outflow, such as Henan, Anhui, Jiangxi, etc. Most of its regional industries are labor-intensive transferred in the eastern regions, such as the first and second categories, so the corresponding role of human capital is not apparent.

Moreover, the results also prove that in the study of Zhang [36], there is a "crowding effect" in the human capital mechanism of urban floating population affecting enterprise export. There is also the same effect in exploring the human capital mechanism of the impact of floating population on enterprise innovation. Thus, when the government implements the recruitment policy, it should not blindly pursue upgrading the human capital structure but pay attention to optimizing it. When implementing the talent introduction policy, all regions should introduce top talents and emphasize cultivating and attracting high-level technical and skilled talents to form a reasonable employment structure and improve the industrial talent system.

5. Conclusion and Countermeasure

This paper comprehensively examines the relationship between multi-dimensional floating population characteristics and enterprise innovation by using the LightGBM algorithm for the first time. Among them, this research summarizes the high-dimensional characteristic variables of the floating population from related studies and studies whether the characteristics of the floating population can predict enterprise innovation to a large extent. Moreover, it is further analyzed the characteristics and prediction model of the floating population with a high degree of prediction of enterprise innovation. It is found that: (1) On the whole, the prediction ability to float population characteristics for enterprise innovation is weak. That is, the maximum lifting effect is 3.031%. (2) Among many floating population characteristics, LnCPCM and PFS features ranked fifth and seventh, respectively. population mobility scale and the proportion of the inter-provincial floating population have a higher ability to predict enterprise innovation. (3) Through the visual analysis of the SHAP model, the relationship between the scale of population flow and the proportion of inter-provincial floating population and enterprise innovation presents a nonlinear characteristic different from the U-shaped relationship in previous studies. Based on analyzing the impact of the interaction of characteristics on enterprise innovation, it is found that after the population flow scale is 35% and the logarithm of the proportion of the inter-provincial floating population reaches 4.4, the model trend shows a population crowding effect, which is consistent with previous theories. (4) It is tested that the population crowding effect has an impact on enterprise innovation through mechanisms o such as human capital, which is consistent with previous research.

This research opens a new dimension of the research on the characteristics of floating population and enterprise development in China: the ability to float population characteristics to predict enterprise innovation. It is of great significance for an in-depth understanding of the role of the floating population in enterprise development. Based on machine learning methods and predictive models, there is still a broad research space to explore in the future. First of all, Scholars can use machine learning methods combined with non-structural data to construct variables that cannot be directly observed to reveal the important characteristics of the floating population that have not been proved in theory. Second, similar models in this study can explore the ability to float population characteristics to predict other indicators of enterprises.

References

[1] Li G S, "The study on the impact of labor cost rising on the industrial technology progress," Huazhong University of Science and Technology, 2015.

[2] Hu Y, Li X Y, Ma J, "A study on the relationship between the upgrading of Human Capital structure and Innovation ability, A case study of floating population in Hebei Province," Modernization of Management, vol. 41, no. 02, pp. 77 - 81, 2021.

[3] Cui T T, Chen X, "Population mobility and urban innovation capability: Evidence from Chinese sub-provincial cities and municipalities," Scientific Management Research, vol. 41, no. 11, pp. 23 - 30, 2021.

[4] G Shmueli, "To Explain or to Predict? ," Statistical, vol. 25, no. 03, pp. 289 - 310, 2010.

[5] Yu X, Chen S K, "Analysis on the difference between inter-provincial population mobility and human capital mobility in China," Population Journal, vol 42, no. 01, pp. 30 - 41, 2020.

[6] Dai H Y, Shi X Z and Lu Y, "labor mobility and corporate innovation," China Journal of Economics, vol. 8, no. 01, pp. 159 – 188, 2021.

[7] Zhong R, "Spatial transfer and Planning strategies of urban-rural aging population reversed in China: from the perspective of population mobility," Urban Development Studies, vol. 26, no. 02, pp. 24 - 30, 2019.

[8] Gu H Y, Meng X, Shen D Y, Cui N N, "Spatial variation of the determinants of China' s urban floating population' s settlement intention," Acta Geographica Sinica, vol. 75; no. 02, pp. 240 - 254, 2020.

[9] Chen D F, Yan Z F, Wang W P, "Urban population size, Industrial agglomeration Model and Urban Innovation, Empirical evidence from 271 cities at prefecture level and above," Chinese Journal of Population Science, no.05, pp. 27 - 40+126, 2020.

[10] Zhou X, Cheng L, Wang H, "Is higher the level of technological innovation, better the financial performance? an empirical analysis of patent applications of the listed chinese pharmaceutical company," Journal of Financial Research, no. 08, pp. 166 - 179, 2012.

[11] G DOSI, L MARENGO, C PASQUALI, "How Much Should Society Fuel the Greed of Innovators? On the Relations between Appropriability, Opportunities and Rates of Innovation," The Capitalization of Knowledge. Edward Elgar Publishing, pp. 13300, 2010.

[12] QING Tao, "Human Capital Investment and Enterprise Innovation: Evidence from China's Micro Enterprise Data", Population & Economics, vol. 03, pp. 108 – 127, 2021.

[13] Li Chuntao, Yan Xiuwen, Song Min and Yang Wei "Fintech and Corporate Innovation--Evidence from Chinese NEEQ-Listed Companies", China Industrial Economics, vol. 01, pp. 81 – 98, 2020.

[14] Sui S M, He Z H, "The influence of population ageing on enterprise innovation: an empirical analysis based on the data of census and micro industrial enterprises", Population research, vol. 44, no. 06, pp. 63 – 78, 2020.

[15] Wang Kai, Li Kai and Liu Tao, "Spatial differentiation and governance efficiency of the citizenization of floating population in china ", City Planning Review, vol. 44, no. 06, pp. 22 - 30+112, 2020.

[16] Yuan Xiewei, "The influence of Labor Age on technological Innovation-- an empirical Analysis based on the Panel data of 238 cities", Productivity Research, vol.18, no. 07, pp. 18 - 24+63, 2021.

[17] Wang Xingdong, "Research on the relationship between urbanization, population density and Innovation", Joumal of Guangdong Institute of Public Administration, vol. 29, no. 06, pp. 73 – 81, 2017.

[18] Lu Y, Zhang Y Q, Li B and Zhao Y H, "Managerial individual characteristics and corporate performance: Evidence from a machine learning approach," Journal of Management Sciences in China, vol. 23, no. 02, pp. 120 – 140, 2020.

[19] R. E. Schapire andY. Freund, Boosting: foundations and algorithms. Cambridge, MA: MIT Press, 2012.

[20] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," Ann. Statist., vol. 29, no. 5, Oct. 2001, doi: 10.1214/aos/1013203451.

[21] Fan D X, "Prediction and application of travel mode selection for urban residents using the light gradient boosting machine method," Huazhong University of Science and Technology, 2018.

[22] Zhang Junbo, he Chuan, Yan Jian, Wu Xianyin and Meng Wei, "Discussion on the applicability of XGBoost algorithm based on cross validation in prediction of rockburst intensity classification," Tunnel construction, vol. 40, no. S1, pp. 247 – 253, 2020.

[23] Li H. Statistical Learning Method. Beijing: Tsinghua University Press, pp. 24 – 25, 2019.

[24] Liang S, Diao J W, Xiao B, "Researchon Risk Predictionof P2P Lending Platform," operations research and management science, vol. 30, no. 01, pp. 170 – 176, 2021.

[25] Lv L D, Duan C R, Liu T and Jin Yongai, "An analysis of the change of floating population in China and the implication," South China Population, vol. 33, no. 01, pp. 20 – 29, 2018.

[26] Wang T, Li Y, "How does the peasant workers' employment policy promote employment? A qualitative comparative analysis based on configuration perspective (fsQCA)," Population & Economics, vol. 04, pp. 37 – 50, 2021.

[27] Wu K P, Tian X, "Imperfect information and imbalances in the migrant labor market," China Journal of Economics, vol. 7, no. 02, pp. 194 – 230, 2020.

[28] Shen J H, "From factory workers to gig workers: changes and invariance of migrant workers' employment in the era of platform economy," Journal of Tianjin Administration Institute, vol. 23, no. 05, pp. 87 – 95, 2021.

[29] [29] Wen X Y, "The trap of de-skilling: Guard against the structural risks of the gig economy to the manufacturing industry," Exploration and Free Views, vol. 11, pp. 150 – 159+180, 2020.

[30] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," 31st Conference on Neural Information Processing Systems, pp. 1 – 10, 2017.

[31] Cheng H, Tang T, "The impact of Labor costs rising on the enterprise innovation of different sizes: Empirical evidence from CEES," Science & Technology Progress and Policy, vol. 33, no. 23, pp. 70 – 75, 2016.

[32] Yang J G, "Study on the influence of urban scale on thetechnological innovation capability, A case study of Yangtze River delta,", Chongqing Technology and Business University, 2017.

[33] Liu Y, Xu X F, Ma H T, "Impact of human capital stock and population concentration on innovative output in China," Scientia Geographica Sinica, vol. 41, no. 06, pp. 923 – 932, 2021.

[34] Wang J W, Liu N N, Sang W C, "Research on the organizational model and structure of urban collaborative innovation development: Taking Hangzhou as an Example," Urban Development Studies, vol. 28, no. 08, pp. 77 – 84, 2021.

[35] Wang D and Zhou Y X, "Application of Random Forest Model in Macro Prudential Regulation-An Empirical Study Based on the Data of 18 Countries," Studies of International Finance, vol. 11, pp. 45 – 54, 2020.

[36] Zhang M Z, Chen R J, "How does floating population affect firms' exports? ," Economic Perspectives, no. 05, pp. 85 – 100, 2020.

[37] He B L, Fan X R, "Comprehensive Evaluation of Independent Innovation capability of large and medium-sized Industrial Enterprises in Chongqing, Shaanxi and Sichuan," Journal of Industrial Technological Economics, vol. 04, pp. 121 – 123, 2008.