International Core Journal of Engineering Volume 6 Issue 12, 2020
ISSN: 2414-1895 DOI: 10.6919/ICJE.202012_6(12).0028

Research on the Prediction of Mass Entrepreneurship

Dongkang Luo

School of Management, Shanghai University, Shanghai 201800, China.

Abstract

China's mass entrepreneurship is conducive to building a new driving force for economic development. Based on the matching data of China's Household Finance Survey and China's urban statistical yearbook, this paper uses support vector machine, K-nearest neighbour algorithm, logistic regression algorithm and decision tree algorithm to predict whether an individual will start a business. Finally, the experimental results of different algorithms are compared and analyzed.

Keywords

Entrepreneurship; Machine Learning; Confusion matrix; Decision tree.

1. Introduction

There are two important aspects of China's economic development in the past 40 years of reform and opening up: the first is opening up to develop China's economy; the second is innovation and entrepreneurship, including cultivating the endogenous growth and activating the entrepreneurial vitality of domestic economic entities. In recent years, one of the government's focuses is to promote the mass entrepreneurship and enhance the participation rate and success rate of entrepreneurship.

It is not easy to start a business. Successful entrepreneurship needs the support of government departments, including create a good business environment and provide personalized services. Therefore, how to accurately identify the people who may start businesses has become an important issue. When the target population is accurately identified, the government can take the initiative to formulate business plans and create good entrepreneurial conditions for them, so as to encourage the public to mass entrepreneurship to the maximum extent.

There are many factors that can affect entrepreneurship. Previous studies have shown that entrepreneurship is the result of the combination of internal individual factors and external environmental factors [1, 2]. Because individual factors are relatively stable in a period of time, entrepreneurial decision-making is more vulnerable to environmental factors such as potential opportunities, entrepreneurial resources and competition. Therefore, urban level factors and individual level factors are both import to entrepreneurship.

China is the preferred country for foreign investment. Enterprises with foreign direct innvestment (FDI) usually have higher technological innovation ability, management level and internationalization degree. Previous studies have also found that FDI can significantly improve the regional economic environment, increase import and export trade, enhance total factor productivity and innovation ability [3]. Therefore, when studying mass entrepreneurship, FDI has become an increasingly important factor.

It is difficult to use mathematical model to measure whether an individual starts a business or not. So this paper uses machine learning to predict it.

The contents of the paper are as follows: the second part introduces the selected indicators; the third part uses support vector machine, random forest regression and logistic regression to carry out experiments, and results are analyzed; the fourth part summarizes the full text.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202012_6(12).0028

2. Indicator Selection

2.1 Data Source

In this paper, the matching data of urban level and individual level are used to predict whether individuals start their own businesses. The data at the urban level are from “China Urban Statistical Yearbook”. The individual level data are mainly from the “China Household Finance Survey” project. The project is carried out by China family finance investigation and research center of Southwest University of Finance and economics. The information of individual characteristics, occupational income and family finance of residents are collected by scientific sampling [4]. Referring to the practice of Yin et al. and Li et al. this paper identifies individual entrepreneurship by whether the individual occupation belongs to "operating an individual or private enterprise, and starting a business independently", and identifies the family entrepreneurship by answering "whether the family is engaged in industrial and commercial operation projects". In this paper, individual entrepreneurship and family entrepreneurship are collectively referred to as mass entrepreneurshipp [5, 6].

2.2 Indicator Selection

Referring to the research of Zhang[7], the indicators are divided into urban level indicators and individual level indicators.

12 urban level indicators are selected, mainly including economic development level (GDP), fixed asset investment level (Invest), average wage level (Wage), proportion of state-owned economy (Share) and FDI.

10 individual level indicators are selected, mainly including age (Age), risk preference (Risk), gender (Gender),whether they own a house (House), whether they are married (Married).

2.3 Descriptive Statistics

There are about 20000 pieces of data. The descriptive statistics of the main indicators are shown in Table 1.

Table 1 Descriptive statistics

	mean	std	min	max
LnFDI	10.622663	2.004789	5.150397	13.993860
Invest	9.906161	0.688643	8.281915	11.286600
GDP	16.917452	1.083855	14.956730	19.072780
Wage	10.530019	0.313644	10.044740	11.251970
Share	0.164207	0.161661	0.000571	0.935189
Age	39.118274	15.029955	4.000000	111.000000
Risk	0.316946	0.201193	0.111111	1.000000
Male	0.554455	0.497039	0.000000	1.000000
Marry	0.734221	0.441759	0.000000	1.000000
House	0.863863	0.342944	0.000000	1.000000
Entrepreneurship	0.177439	0.382050	0.000000	1.000000

ISSN: 2414-1895 DOI: 10.6919/ICJE.202012_6(12).0028

3. Experiment and result ananlysis

In this section, support vector machine, K-nearest neighbour, logistic regression and decision tree are used to predict whether an individual starts a business or not, and the results are compared and analyzed.

3.1 Confusion matrix

Confusion matrix as shown in Table 2 is built.

Table 2 Confusion matrix

		Predicted	
		+1	0
Actual	+1	TP	FN
	0	FP	TN

An individual starting a business is defined as positive, otherwise negtive. The main considerationn of this study is the individual starting a business, so recall is selected to evaluate the result other than accuracy. The formula of recall is shown as below:

$$\boldsymbol{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP + FP}}$$

3.2 Support Vector Machine Algorithm

Support vector machine (SVM) algorithm is a generalized linear classifier which classifies data according to supervised learning. Its decision boundary is the maximum margin hyperplane. The main feature of SVM is that it can construct the decision boundary with the maximum distance, so as to improve the robustness of classification algorithm. Therefore, SVM can avoid over fitting problem, and the effect is relatively better when the data is insufficient.

Common kernel functions include linear function, Gaussian kernel function and polynomial kernel function. The polynomial kernel function can fit the hyperplane with complex separation, but it is difficult to choose its parameters because of its large number of parameters. Gaussian kernel function can map the input features to infinite dimensions, and the calculation is moderate, but it is easy to cause over fitting and the result is not intuitive. The linear kernel function is simple and efficient, but it can not deal with complex problems. To sum up, Gaussian kernel function are choosed as the kernel function of SVM.

The result is shown in Table 3.

Table 3 Result of SVM

		Predicted	
		+1	0
Actual	+1	3039	0
	0	131	525

The recall is 80.03%, and the accuracy is 96.45%.

3.3 K-nearest Neighbour Algorithm

The principle of K-nearest neighbour(KNN) algorithm is to calculate the distance between the data sample to be marked and each sample in the data set, and take the nearest K samples. The category of the data sample to be marked is generated by the voting of the k nearest samples. Its advantages are high accuracy and high tolerance to outliers and noises. The disadvantage is that the amount of calculation is large.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202012_6(12).0028

Through comparison, choose $K = 3$. The result is shown in Table 4.

Table 4 Result of KNN

		Predicted	
		+1	0
Actual	+1	3013	26
	0	385	271

The recall is 51.37%, and the accuracy is 88.88%.

3.4 Logistic Regression Algorithm

Logistic regression(LR) algorithm is a classification algorithm. Generally, sigmoid function is used as prediction function to achieve classification purpose. Sigmoid function is as follows:

$$g(\theta^T x) = \frac{1}{1 + e^{\theta^T x}}$$

Sigmoid function has the property that the value of function changes significantly near $x = 0$, and it approaches to 1 or - 1 infinitely when x is large enough or small enough, respectively.

The result is shown in Table 5.

Table 5 Result of LR

		Predicted	
		+1	0
Actual	+1	3039	0
	0	28	628

The recall is 95.73%, and the accuracy is 99.24%.

3.5 Decision Tree Algorithm

Decision tree(DT) algorithm is a method to approximate the value of discrete function. It is a typical classification method. In essence, decision tree is a process of data classification through a series of rules. Its main advantage is that the model is easy to explain and the classification speed is fast.

The result is shown in Table 6.

Table 6 Result of DT

		Predicted	
		+1	0
Actual	+1	3018	22
	0	25	631

The recall is 96.19%, and the accuracy is 98.76%.

3.6 Result Analysis

The results of all algorithms is shown in Table 7.

Tabel 7 Results of all algoridothms

	SVM	KNN	LR	DT
Recall	80.03%	51.37%	95.73%	96.19%
Accuracy	96.45%	88.88%	99.24%	98.76%

ISSN: 2414-1895 DOI: 10.6919/ICJE.202012_6(12).0028

As shown in table 7, When decision tree algorithm is used, the recall rate is the highest, and its accuracy is close to that of logistic regression algorithm, whose accuracy is the highest. Therefore, decision tree algorithm has the best performance in predicting whether an individual starts a business or not.

4. Conclusion

This paper takes the data from China Urban Statistical Yearbook and China Household Finance Survey as samples to predict whether an individual will start a business by several algorithms of machine learning, including support vector machine, K-nearest neighbour algorithm, logistic regression algorithm and decision tree algorithm.

References

[1] Shane, S., E. A. Locke, and C. J. Collins. Entrepreneurial Motivation [J]. Human Resource Management Review, 2003, 13(2): 257-279.

[2] Taormina, R. J., and S. Kin -Mei Lao. Measuring Chinese Entrepreneurial Motivation: Personality and Environmental Influences [J]. International Journal of Entrepreneurial Behavior & Research, 2007, 13(4): 200- 221.

[3] Liu X L, Xiong X. Foreign direct investment, import and export trade and regional economic growth -- a case study of Hunan Province [J]. Management World, 2016, (2), 184-185.

[4] Gan L, Yin Z C, Jia N, et al. Research report of China household finance survey 2012[M]. Chengdu, Chinan, Southwest University of Finance and Economics Press, 2012.

[5] Yin Z C, Song Q Y, Wu Y, et al. Financial knowledge, entrepreneurial decision and entrepreneurial motivation [J]. Management World, 2015, (1), 87-98.

[6] Li X L, Ma S, Lv R S. Spillover effects of multinational corporations in China: from the perspective of human capital [J]. Economic Research Journal, 2015, (5): 89-103.

[7] Zhang K D, Wu X F, Gao J, et al. The impact of foreign direct investment on mass entrepreneurship[J]. China Industrial Economics, 2018, 369(12):81-98.