Based on SPSS Analysis of Lending Factors on Small Bank Enterprises Dataset

Xinling Zhao[1,a], Zhitao Liu[2,b*], Sicheng Chen[3,c]

[a]cytheriazhao@gmail.com, [b]yamin2024@163.com, [c]Seasonnnchennnn@qq.com
* Corresponding author. Email: yamin2024@163.com

[1]College of Accounting of Guangzhou Institute of Business and Technology, Guangzhou College of Technology and Business, Guangzhou,528138, China
[2]School of Computer of Science, Zhuhai College of Jilin University, Zhuhai, 519041, China
[3]School of Sedbergh, Fuzhou, 350207, China

Abstract. As the society is currently in a recession and loans have become the center of attention, the problem of loan defaults is also a major concern for lending institutions. In this paper, we use the Small Business Administration dataset to analyze the data to help credit institutions such as banks to issue loans more efficiently and accurately in order to reduce the amount of bad debts and the amount of guarantors' burden. Specifically, using chi-square tests, we found that old loans have significantly higher loan amounts and repayment rates compared to new loans, and that having a real estate guarantee significantly increases their loan amounts and repayment rates. We also conducted a factor analysis and found that total payments, loan amount, and guarantees for small bank businesses serve as the first principal component, i.e., for loans can provide the most effective amount of information. In addition, the results of ANOVA showed that as the number of lending operations in a state increases, total loans, small bank business guarantees, and total payments significantly increase. Our study is useful in helping credit institutions such as banks to provide more efficient and accurate loan disbursements in order to reduce bad loans and SBA guarantees.Keywords: Loan, Independent sample t-test, Chi-Squared Test, Cross Tabulations, Singer Factor ANOVA, Factor Analysis.

Keywords: Loan, Independent sample t-test, Chi-Squared Test, Cross Tabulations, Singer Factor ANOVA, Factor Analysis.

1. Introduction

1.1. Backgound

There are success stories of startups such as FedEx and Apple Computer receiving Small Business Administration loan guarantees. In other countries where the traditional banking sector is dominant and the main income of the banks is a large percentage of bank assets and interest income on bank loans, an inappropriate assessment of the loan portfolio may also have a negative impact on the performance of the commercial banks and even on the economic increase of the country and the banking system. However, some small businesses or startups may default on SBA-guaranteed loan stories. There has been controversy over the default rate on these loans for decades. Some conservative economists argue that the credit market is able to regulate itself without government involvement, while proponents of SBA-guaranteed loans argue that the

ICEMBDA 2023, October 27-29, Tianjin, People's Republic of China

DOI 10.4108/eai.27-10-2023.2341934

small business when obtaining government guaranteed loans, the societal benefits of job creation regarding jobs greatly outweighs the cost of defaulting on the loan. In this work,in order to find out more reasons and continue to solve this problem. We use the Small Business Administration dataset to analyze the data to help banks and other credit institutions make loans more efficiently and accurately and reduce bad debts and guarantor liabilities.

1.2. Dataset Background

The background of this dataset is that Small business administration established in 1953 to help improve and assist in the development of the small business credit market in the United States. The main source of jobs in the United States is the creation of small businesses. Therefore, it is more socially beneficial to help small business creation and development in order to create jobs and stabilize the unemployment rate. One way that the Small Business Administration helps small businesses is through the Small Business Loan Guarantee Program, which is designed to help small businesses get better loans from banks. The SBA is like an insurance company that guarantees a portion of a loan to reduce the bank's risk. If the loan defaults, the SBA pays for the guaranteed amount. The rest of the paper is organized as follows.In Section 2, we explain the experimental analysis methodology. In Section 3, we explain the experimental results of the analytical approach. In section 4, a one-way anova analysis will be conducted. Section 6 will conclude with a summary.

2. Experimental analysis methods

2.1. Chi-Square Test

The chi-square test is a statistical method proposed by Pearson, which compares the probability value of the chi-square distribution function with the chi-square statistic, and determines whether the expected probability is consistent with the actual probability under a certain level of confidence and degree of freedom, and the correlation of two categorical variables can be analyzed by comparing the degree of similarity between the theoretical probability and the actual probability. Chi-square test is used to find if there is any correlation among nonnumeric variables that are frequently used in statistical studies [1, 2].

Users can use SPSS software to easily complete the chi-square test, in SPSS software, we default H0 is success, that is, there is no difference between the observed frequency and the real frequency, that is, the two groups of variables will not affect each other, and there is no correlation between the two groups of variables. If the test P value is very high, the hypothesis test passes; if the test P value is very low, the test fails, and the difference between the observed frequency and the real frequency of the two groups of variables is correlated. SPSS data tests are highly scientific and complete, and therefore give reports that are relatively difficult to analyze.

2.2. Independent Samples t-test

The independent samples t-test, known as the group t-test or two independent samples t-test, is often used to compare the means of two samples in a completely randomized experiment In medical analysis, assigned to two different treatment groups in a completely randomized fashion, the analyst focuses primarily on whether the two overall means will be equal in the mean of the

two samples [3]. In addition, in exploratory studies, independent sample t-tests are also often used for mean comparisons in two fully machine-sampled samples.

2.3. One-way ANOVA

There are some underlying assumptions of one-way analysis of variance, and the first one is: Each individual sampling unit selects its own reference group independent of all others (sampling independence). In other words, there is no correlation between the samples, such as picking samples from the same person, but at different time periods; there are different tests for this situation [4].

One-way ANOVA is used to analyze whether there is a significant correlation between the means of multiple samples under the influence of a single control variable. This method is used to analyze whether there is a correlation between the means of different levels of the dependent variable under the influence of the control variable. One-way ANOVA assumes that the variance of each observation is equal between subgroups of control variables from a normal sample, and that the subgroups are independent and at different levels. The analysis will categorize all variance into explainable bias (systematic bias) and unexplained random bias. If the random bias is significantly smaller than the systematic bias, then there is significant correlation due to the means when the controls are at different levels [5].

2.4. Factor analysis

Factor analysis is carried out under the idea of dimensionality reduction, to maximize the likelihood of not losing or less loss of information on the original data To be aggregated into a small number of public factors as much as possible, aggregated from an intricate multitude of variables, while reducing the number of variables, the information of many original variables can still be reflected, and can show the intrinsic correlation between variables[6].Usually,factor analysis is generally used for three purposes: one is used for factor dimensionality reduction, the second is to calculate the factor weights, and the third is to calculate the weighted factor summary composite score.

Factor rotation is a case of matrix linear transformation. Visual rotation was the original method and is still used as a standard that many analytical methods attempt to replicate; Visual rotation can now rival analytical rotation from an efficiency standpoint by programming appropriate procedures for time-sharing computer systems [7].

3. Comparative analysis

3.1. Chi-Square Test

3.1.1. NewExist Repayment rate analysis

Table 1.MIS_Statues*NexExist crosstab.

NewExist					
			1	2	Total
MIS_Statues	CHGOFF	count	110114	47367	157481

		Percentage of MIS_Status	69.9%	30.1%	100.0%
		Percentage of NewExist	17.1%	18.7%	17.5%
	P I F	count	533332	205192	738524
		Percentage of MIS_Status	72.2%	27.8%	100.0%
		Percentage of NewExist	82.7%	81.1%	82.2%
	Total	count	644869	253125	897994
		Percentage of MIS_Status	71.8%	28.2%	100.0%
		Percentage of NewExist	100.0%	100.0%	100.0%

There is no big difference between the repayment rate of old business and new business (Chi-Squared Test, Cross Tabulations). According to Table 1, it can be seen that the business with more than two years of history has a stable repayment rate of 82.7, while the new business is 81.1, and the outstanding old business is 17.1, while the new business is 18.7, which can be seen that it does not have a significant difference. Through the chi-square test to determine whether it is also because the old business has a more stable repayment rate than the new business, because the new business and the old business may be based on the strength of their respective enterprises to the bank loan, so the repayment rate should also be similar.

We use machine learning a lot when producing labeled or unlabeled profiles for training models; it's also helpful in spotting potential patterns and trends. Models using machine learning can calculate the likelihood of a particular employee leaving, which can be used by senior managers to take the initiative.[8].

3.1.2. Realestate and MIS_Statues analysis

Table 2.MIS_Statues*Realestate crosstab.

Realestate					
			0	1	Total
MIS_Statues	CHGOFF	count	155081	2477	157558
		Percentage of MIS_Status	98.4%	1.6%	100.0%
		Percentage of Realestate	20.8%	1.6%	17.5%
	P I F	count	589978	149631	739609
		Percentage of MIS_Status	79.8%	20.2%	100.0%
		Percentage of Realestate	79.0%	98.3%	82.3%
	Total	count	747007	152157	899164
		Percentage of MIS_Status	83.1%	16.9%	100.0%
		Percentage of NewExist	100.0%	100.0%	100.0%

Firms secured by real estate have higher repayment rates than firms not secured by real estate (Chi-Squared Test, Cross Tabulations).

According to Table 2, Loans collateralized by real estate have a lower default rate of 1.6% compared to other collateral, while loans collateralized by non-real estate are higher than 20% (risk of default). Research has shown that in recent years an increasing number of problem loans have been made by developments in employment rates and real interest rates [9].

The full payment rate for loans secured by real estate is 98.3%, while the full payment rate for loans not secured by real estate is 79.0%.

In order to see whether the real estate guarantee has a significant impact on bank loans, from the cross-tabulation table can be seen that most of the enterprises do not have real estate guarantee, no real estate guarantee accounted for 83.1 percent, with real estate guarantee of only 16.9 percent, it can be seen that the real estate guarantee is not part of the enterprise can not be done, or most of the enterprises need loans do not need real estate guarantee, and secondly, it can be seen that no Secondly, it can be seen that the repayment rate of enterprises without real estate guarantee is only 79 percent, while the repayment rate of enterprises with real estate guarantee is close to 100 percent at 98.3 percent, which can be further seen that real estate guarantee has a significant impact on repayment, and it has a positive and active impact on the promotion of repayment.

3.1.3. Impact of repayment rates and number of loans in rural municipalities

Table 3.MIS_Statues*Realestate crosstab.

		UrbanRural	
		city	Rural
MIS_Status	CHGOFF	114867	19713
	P I F	354414	85347
	Total	470654	105343

The number of loans is higher in urban than in rural areas, but the repayment rate is higher in rural than in urban areas. According to Table 3, it can be seen that the loan rate of the city is more than 3 times more than that of the rural area.

Table 4.MIS_Statues*Realestate crosstab.

			Realestate		
			1	2	Total
MIS_Statues	CHGOFF	count	114867	19713	134580
		expectation count	109966.9	24613.1	134580.0
		Percentage of MIS_Status	24.4%	18.7%	23.4%
	P I F	count	354414	85347	439761
		expectation count	359333.9	80427.1	439761.0
		Percentage of MIS_Status	75.3%	81.0%	76.3%
	Total	count	470654	105343	575997
		expectation count	470654.0	105343.0	575997
		Percentage of MIS_Status	100.0%	100.0%	100.0%

According to Table 4, the repayment rate of the rural area is 81.0%, while the city is 75.3%, it can be seen that the repayment rate of the city is not as good as the rural area, because of the developed economy in the city, there are many enterprises, so the number of loans is more, but because of the pressure of survival, the economy is even more difficult to support, resulting in the repayment rate is also lower than the rural area.

Lending for residential real estate (RRE) is a major activity for euro area banks, so that market can be considered of great importance from both a microprudential and a systemic risk perspective, given its ties with the real economy and the financial sector [10].

The repayment rate is also lower than that in rural areas, and secondly, loans are given to rural areas because rural enterprises are generally more stable, and the economic pressure is lower than that in cities, and the repayment rate is higher than that in cities.

Table 5. chi-square test.

	value	Asymptote salience (bilateral)
Pearson's chi-square	1563.799	.000
logarithmic ratio	1627.079	.000
Number of active cases	575997	

According to Table 5, the significance is 0.000, so there is a significant difference between urban and rural loans.

3.2. Independent samples t-test

3.2.1. NewExist And SBA_Appr Analysis

Table 6. Group statistics.

	NewExist	Number of cases	Average value
SBA_Appv	Old	644869	158305.4803
	New	253125	126529.0593

Table 7. Independent samples test.

		F	Significance	Significance (two-tailed)
SBA_Appv	Assuming equal variance	7481.864	.000	.000
	Not assuming equal variance			.000

The total amount of guarantees for small bank businesses is much larger for old businesses than for new businesses in terms of both the number of guarantees and the average value of guarantees (Independent sample t-test). Explore whether the new business has any effect on the amount of small business bank loans, 1 for business more than 2 years old and 2 for business less than or equal to 2 years old.

According to Table 6 can be seen 158305 and 126520, the old business than the new business 3 w, followed by the number of guarantees the old business has 644869 while the new business for 253125, the old business than the new business more than the gap of nearly 400,000, According to Table 7 Levin variance equivalence test for 0.00 is less than 0.05, proving that there is a significant difference, looking at not assuming equivalence of variances, it can also be seen that the significance 0.00 < 0.05, so proving that there is a significant difference, and from the group statistics. Further proved out that the emerging business and old business also have a more significant impact on the bank to grant loans. We can see that the old and new business

has a huge impact on the guarantee of small bank enterprises, because the old business can be a long and stable operation, while the new business has too much uncertainty and more risk.

3.2.2. NewExist And GrAppr Analysis

Table.8 Group statistics.

	NewExist	Number of cases	Average value
GrAppv	Old	644869	204122.42
	New	253125	162987.10

Table 9. Independent samples test.

		F	Significance	Significance (two-tailed)
GrAppy	Assuming equal variance	6230.845	.000	.000
	Not assuming equal variance			.000

Old business has a larger mean total guarantee and larger number of loans than new business (Chi-Squared Test, Cross Tabulations).

1=existing, 2=new. Whether the firm is a new or established business (denoted as "NewExist" in the dataset, with "Newexist" = 1 if the business is less than or equal to 2 years old, and "Newexist" = 1 if the business is more than 2 years old as an established business). (If the business is more than 2 years old, then "Newexist" = 0).

According to Table 8 it can also be seen that the total loan amount of old and new business has a difference of nearly 4w, which can be further seen that the bank's amount for old and new business is also significantly different. Moreover, the number of cases is 644869 and 253125 has a difference of nearly 40w.

According to Table 9, it is argued that the failure rate of new business is likely to be higher than that of the existing old business. Established businesses are mature enough to have achieved economies of scale, their business processes are mature and cheaper than new businesses, and they (established businesses with successful operations) are applying for loans to expand their mature and already successful existing businesses with a higher chance of continued success. However, new businesses may not be able to anticipate the obstacles they face and are unable to fully circumvent as well as successfully overcome them, which in turn leads to loan defaults. Therefore both the number of loans and the average value are smaller than the old business.

However, when comparing the loan default rates of new business (less than or equal to 2 years old) and established business (more than 2 years old) in this dataset, the difference between the two is relatively negligible.

3.2.3. Realestate And GrAppr

Table.10 Group statistics.

	Realestate	Number of cases	Average value
GrAppv	Not	747007	134926.6339
	Have	152157	476258.4202

Table.11 Independent samples test.

		F	Significance	Significance(two-tailed)
GrAppy	Assuming equal variance	98933.538	.000	.000
	Not assuming equal variance			.000

Property guarantees result in larger loan amounts than no property guarantees (Independent sample t-test).

According to Table 10, it can be seen that the amount of loans secured by property is three times the amount of loans not secured by property, which further proves that most of the loans secured by property are large loans, and the banks are more comfortable in lending to those who have large loans.

According to Table 11 and 12 the significance of 0.00 can be seen from the significance of 0.00 in the Unassumed Equal Variance that there is a significant difference.

Therefore whether the loan is secured by real estate (land ownership) serves as another indicator of risk. The rationale for this indicator is that the value of the land is large enough to cover the amount of any outstanding principal and thus can reduce the probability of default.

4. One-way ANOVA

4.1. State and GrAppr and SBA_Appr and DisbursementGross

Table 12. Describe.

		Number of cases	Average value
GrAppy	1.0	130619	263718.06
	2.0	70458	217590.59
	3.0	57693	134155.60
	4.0	41212	198424.63
	5.0	35170	139707.99
	6.0	32622	157077.58
	Total	367774	205921.62
SBA_Appr	1.0	130619	207836.95
	2.0	70458	165358.16
	3.0	57693	99254.55
	4.0	41212	152805.40
	5.0	35170	102796.51
	6.0	32622	119695.25
	Total	367774	158635.52
DissbursementGross	1.0	130619	269304.78
	2.0	70458	223253.74
	3.0	57693	147857.26
	4.0	41212	206791.72
	5.0	35170	151040.45
	6.0	32622	168487.01
	Total	367774	214173.48

There is a significant positive correlation between total loans, guarantees to small bank businesses, and total disbursements and the number of cases in a state (Singer Factor ANOVA). We conducted an ANOVA analysis in order to explore whether there is significant variability between states, taking the six states with the largest amount of data.According to Table12 it can be seen that the more businesses the bank lends to, i.e., the higher the number of cases, the more the SBA_Appv and GrAppv and DisbursementGross averages will also increase significantly. It is also in line with the law of reality that usually the more loans are made to banks, the more businesses there are in the state, the more developed the state's economy is, the larger the business, the larger the amount of loans needed, and naturally the more the small business bank dares to guarantee. Secondly, because of the law of economic development, most economically developed areas, will become more and more developed, the more the bank also believe that the enterprise has a higher repayment ability, the more the loan amount.

4.2. Relationship between Urbanrural and SBA_Appr Analysis

Table.13 Describe.

		Number of cases	Average value
GrAppy	City	470654	.6408
	Rural	105343	.6852
	Total	575997	.6489

Table.14 ANOVA.

percentage	square sum	Degrees of freedom	mean square	F	significance
intergroup	169.576	1	169.576	5752.253	.000
Within a group	16980.310	575995	.029		

There is no significant difference in the amount of guarantee for small banking enterprises in rural and urban areas (One Way ANOVA). According to Table 13 and 14 can be seen in rural and urban small bank enterprises guarantee amount does not have a significant difference, can be seen for small bank enterprises in rural and urban does not have a big difference, and ANOVA F is much greater than 0.05, proving that the original hypothesis is valid. And this is also in line with reality, for the rural areas instead of the bank is more willing to guarantee, and there are policy support, while for the city and instead of a little lower, because the risk of the city is higher.

5. Factor analysis

Table 15. Correlation Matrix.

		SBA_Appr	GrAppr	DisbursementGross	NoEmp	Term	ChgOffPrinGr
relevance	SBA_Appr	1.000	.974	.940	.093	.525	.165
	GrAppr	.974	1.000	.971	.090	.503	.194
	DisbursementGross	.940	.971	1.000	.089	.466	.192
	NoEmp	.093	.090	.089	1.000	.046	.008
	Term	.525	.503	.466	.046	1.000	-.047
	ChgOffPrinGr	.165	.194	.192	.008	-.047	1.000

We further explore, what is having more information in this dataset, according to Table 15, SBA_Appv and GrAppv and DisbursementGross all three correlations are greater than 0.9, which are significantly correlated, followed by Term as the next correlation, and Term and ChgOffPrinGr are the least of the six variables in this correlation of the six variables.

Table 16.KMO and Bartlett's test.

KMO Sample Fit Volume	.761	
Bartlett's test of sphericity	Approximate chi-square	5649095.192
	Degrees of freedom	15
	Significance	.000

According to Table 16, the KMO and Bartlett's test in which the KMO sampling aptitude quantity is 0.761 > 0.6 proves that it is possible to perform factor analysis.

Table 17. Total Variance Explained.

	Initial eigenvalue			Extract the sum of the squares of the intercepted loads		
Ingredient	Total	Percentage of variance	Cumulative %	Total	Percentage of variance	Cumulative %
1	3.300	55.003	55.003	3.300	55.003	55.003
2	1.036	17.259	72.262	1.036	17.259	72.262
3	.988	16.470	88.731	.988	16.470	88.731
4	.603	10.046	98.777			
5	.057	.956	99.733			
6	.016	.267	100.000			

According to Table 17, it can be seen that the first three principal components contain the most amount of information, and the amount of information plummets when four, so the first three are taken as the principal components.

Table.18 correlation matrix.

		SBA_Appr	GrAppr	DisbursementGross	NoEmp	Term	ChgOffPrinGr
relevance	SBA_Appr	1.000	.974	.940	.093	.525	.165
	GrAppr	.974	1.000	.971	.090	.503	.194
	DisbursementGross	.940	.971	1.000	.089	.466	.192
	NoEmp	.093	.090	.089	1.000	.046	.008
	Term	.525	.503	.466	.046	1.000	-.047
	ChgOffPrinGr	.165	.194	.192	.008	-.047	1.000

According to Table 18, SBA_Appr and GrAppr and DisbursementGross are the first principal components containing the most information, while ChgOffprinGr is the second principal component and NoEmp is the third principal component containing the least information.

6. Conclusion

We analyze the dataset to show that urban-rural, loan tenure, property guarantees, and small bank business guarantees are of great importance to banks for lending. In addition, the data related to lending varies from state to state. This thesis experiment is important to analyze the bank lending in detail.

REFERENCES

[1] Mačerinskiene I, Ivaškevičiūte L. The evaluation model of a commercial bank loan portfolio[J]. Journal of Business Economics and Management, 2008, 9(4): 269-277.

[2] Turhan N S. Karl Pearson's Chi-Square Tests[J]. Educational Research and Reviews, 2020, 16(9): 575-580.

[3] Eyduran E, Duman H. Application of independent sample t-test and normality tests in R-Lecture notes[J]. 2019.

[4] Connelly L M. Introduction to analysis of variance (ANOVA)[J]. Medsurg Nursing, 2021, 30(3): 218-158.

[5] Kulkarni H V, Patil S M. Uniformly implementable small sample integrated likelihood ratio test for one-way and two-way ANOVA under heteroscedasticity and normality[J]. AStA Advances in Statistical Analysis, 2021, 105(2): 273-305.

[6] Harman H H. Modern factor analysis[M]. University of Chicago press, 1976.

[7] Gorsuch R L. Factor analysis: Classic edition[M]. Routledge, 2014.

[8] Berge T O, Boye K G. An analysis of banks' problem loans[J]. 2007.

[9] Gaudêncio J, Mazany A, Schwarz C. The impact of lending standards on default rates of residential real-estate loans[J]. ECB Occasional Paper, 2019 (220).

[10] Subhashini M, Gopinath R. Employee attrition prediction in industry using machine learning techniques[J].International Journal of Advanced Research in Engineering and Technology, 2020, 11(12): 3329-3341.