Migration Status and Loan Default*

Ruyang Chengan[1] and Jin Xi[1]

[1]Department of Economics, UNC at Chapel Hill

Abstract

The default risk of personal credit loan has been studied a lot over the past few decades. This study examines the effect of immigration status on loan default risks based on data collected from a commercial bank centered in southeast China. It employs logistic regression to model the impact immigration status has on the likelihood of default, but also survival analysis to estimate how immigration status affects the time to default. The result shows that at high income level, immigrants are less likely to default compared to non-immigrants. At low income level, however, the probability of default for immigrants is higher than that of non-immigrants. When income level reaches around 80th percentile, immigrants and non-immigrants have about the same default risk. During the migration process, two groups of people are selected: people with great skills and endowment, and people who mainly engage in manual labour. High mobility and low job security are the features for manual labour, which increase those immigrants likelihood to default when comparing to the locals. The first group, usually with high income, are usually favored by employers over local workers.

*This research is supported by the Herbert Brown Mayo Summer Research Fund. We also want to express our gratitude to Professor Mike Aguilar, who advised this research.

1. Introduction

Personal loan default has been extensively studied since decades ago. Historically, credit scoring systems were built to estimate how likely a borrower will default in a given period of time. Since the late 90s, instead of focusing on whether borrowers will default, researchers ask when will it be if they default. The time of default is crucial because it gives the bank a better view of the profitability. The use of survival analysis on this topic has been articulated in many studies (Banasik et al., 1999). Previous studies of this kind usually employ data of the loan applicants characteristics and the macroeconomic conditions as predictors. However, one important factor has been overlooked the migration status of the applicants.

Since the opening-up reform in China, an upsurge in the movement of human capital was driven by the rapid growth in manufacturing jobs in the urban coastal areas. The volume of migration almost tripled from 12 to 32 million during 1995 to 2000. However, without local hukou (permanent residency permit), most of the workers were temporary migrants, or the so-called floating population, from rural areas (Fan 2007). These migrants in urban areas, however, are treated differently from local workers. Specifically, some of the migrants are especially preferred by employer over local workers, depending on the wage level. A recent study (Wang et al, 2015) compares the treatment in job market of local workers, urban migrants and rural migrants, and finds that urban migrants are preferred to urban locals at high wage level. One explanation is that the prohibitive costs of migration selects out people with strong motivation, ambitions, specialties, and better endowment. Therefore, the migration status in a way is a signal of high qualification, and thus migrants preferred by employers. Rural migrants, on the other hand, are favored over local workers at lower-tier jobs. One explanation proposed in the study is that most local workers tend to shun low-tier jobs, while rural migrants who have little education and skills are willing to do these laborious jobs and earn relatively low wage. However, the preference by employers gives rural migrants little guarantee of secured jobs, because without specialties they can be easily substituted.

Due to the preference of employers, we hypothesize that migrant workers with low-tier jobs are faced with most default risk, because not only are they faced with discrimination as outsiders, but also their jobs by nature are characterized with high mobility. At high wage level, migrant workers get more favors by employers and more secured jobs to a point where migrants might be even less likely to default compared with local workers.

To exam the effect of migration status on loan default risks, this study employs both logistic regression and survival analysis. The logistic model aims to answer what impacts does immigration status have on the likelihood of default, and the survival analysis answers how it affects the time to default. The data used in this study is provided by a commercial bank centered in southeast China.

The remainder of the paper will be organized as follows. We start with introducing the background of migration and loan default in China. Following the overview of our dataset, we proceed to describe the methodology used. Next, we present the test results of the two models, and this is followed by our discussion and conclusions.

2. Background

2.1. Migration in China and The Segmented Labor Market

Implemented since 1950, the hukou system (permanent residency registration) in China used to serve as a strict restriction that prohibits migration from rural to urban areas. Without urban hukou, it was almost impossible for rural migrants to get access to housing and employment in cities (Sun and Fan, 2011). Since the opening-up reform in 1978, a large number of jobs in manufacture were created due to the rapid growth in cities (Fan, 2003). Therefore, to attract cheap labor force, the migration restriction was relaxed considerably to allow rural residents to work temporarily in urban areas, despite they did not own the urban hukou. These temporary migrants make up the so-called “floating population”. From 1989 to 1993, the population of rural migrants surged from 30 million to 62 million (Li, 2008). A report from the National Health and Family Planning Commission also indicates that the population of temporary migrants has reached 236 million in 2012, 75% of which is consists of rural-to-urban migrants.

Although mobility has increased since the economic reform in 1980s, Hukou system is still a major factor leading to the formation of a two-class urban society (Zhu 2007), since migrants without local hukou are treated unequally when applying for jobs and receiving welfare. In the job market, rural migrants are usually considered to be inferior to local workers, as studies show that most of the wage differences cannot be explained by productivity-related characteristics between the two groups (Laurence, 2002). In 1996 a survey conducted in Shanghai also showed that a considerable number of rural migrants were employed in informal or low-wage jobs, because government policies prohibited them from finding good jobs. In terms of social welfare, migrants without local hukou are entitled to limited social welfare unless they pay extra fees (Zhao, 2000, and Meng and Zhang, 2001). The upsurge of housing prices in some big cities has driven the local governments to announce housing purchase limitation that increases the difficulty for migrants to purchase housing. The particular policies include raising the mortgage rates and increase down payment for second home for migrants. For instance in Shanghai, single migrants are not allowed to purchase housing. Because of the unequal treatment, many migrants lived in deprived housing areas (Song et al., 2008).

The reform of state-owned enterprises in late 1990s further induced segmentation between local and migrants. Millions of urban workers lost their jobs during the reform, and they were faced with an increasingly competitive job market because of the surge of migrants. Consequently, migrants were regarded as competitors who made local workers worse off. A series of policy were carried out by local government to protect urban workers. For instance, a 1995 Beijing city document explicitly lists a number of jobs as eligible to only local residents, including managers in finance, accountant, casher, warehouse staff etc. (Bai and Song 2002). These policies were terminated recently, as the central government asked local governments to promote work places with equal opportunities for migrant workers. However, the depth and scope of the reform are improving slowly, as the segmentation has dominated the labor market in the past 20 years (Wang et al. 2015).

2.2. Loan Default in China

In China, credit history is not shared among banks or institutions. All the credit information of an individual or enterprise is collected and only can be assessed from Credit Reference Center of the Peoples Bank of China, known as Credit Bureau. To make decisions on loans, bank need to request reports of credit history from Credit Bureau. The report includes the credit records for past 5 years. Similarly, every time a delinquent or any violation is occurred, the record will be automatically sent to the Credit Bureau.

Banks in China often have strong incentives to lower the interest rate on repayment amount of defaulted loans, owing to the considerable costs of debt collection. Most of the time, delinquent is occurred unintentionally. For example, customers might miss the deadline for repayment by mistake. Usually these type of customers would pay the overdue once they realize. The intentional breach of contract, however, occurs when customer is either not willing to pay or losing the ability to pay. The overdue caused by intentional breach of the contract would cause excessive costs since a complex procedure will be followed.

Once a delinquent occurs, there are several steps for the bank to urge the payment,such as message informing, call collection, civil inquiry, or hire collection agencies in some extreme circumstances. In China, different banks have slightly different procedures for overdue collection. The cost of collecting overdue depends on how fast the customer is able to pay. Particularly for overdue that caused by intentional breach, the procedures might cause an excess burden for the bank. For instance, civil inquiries of default loans usually take over a year to settle. Since excessive costs might occur during the processes, the bank might adjust the interest rate lower to control its cost on delinquent.

The central bank of China also takes action to control the number of default loans. In the past few years, the average loan default rate in China has increased significantly from 0.77% in 2013 to 1.74% in 2017. Consequently, the central bank stepped in with three strategies to compensate for the loss of commercial banks. First, the central bank promotes lower loan interests to mitigate the increase in bad loans. Second, the deposit interest rate was lowered to reduce the costs of bank.[1] Third, each year the central bank buys off some of the bad loans from commercial banks.

3. Data Description

3.1. Overview

The information of personal credit loans is provided by a commercial bank in China[2]. The bank was initially established in 1988. Over the last 10 years, it has expanded rapidly into 12 southern cities. The amount of capital has sized up to 200 billion RMB till 2017. Our dataset covers the information of bank branches in 7 major cities. A brief overview of the cities and their sub-branches is shown in Table 1 and 2.

The 7 cities vary substantially in their industrial structure. Traditionally in China, industries are classified into 3 major categories. The first category includes agriculture, forestry, livestock production, and fishing. The second category includes mining, manufacturing, construction, production of electric power, heat, gas, and water. The third category includes industries other than the first two categories, and it is mainly characterized by service jobs. Figure 1 presents the average GDP in each city generated from the 3 categories throughout 2013 to 2016.

3.2. Characteristics of Loans

The dataset includes application information of over 56,269 loans accepted from August 2013 to February 2017 together with their default status. Due to privacy reasons, activities of borrowers are not tracked, and thus we can not control for borrowers who took out more than one loans over the sampled period. Therefore we treat each loan as a separate entity.

[1] footnote: both fall in the loan interest and deposit rate, we cant say anything about the spread between the two

[2] Due to confidencial reason, we are unable to provide the name of the bank

Information about each loan includes the starting month, time limit of the contract, loan amount, fixed annual interest rate, type of the loan, and the city where the loan was originated. Table 3 presents a brief summary of basic statistics.

In the data set, all loans are based on credit guaranty, which is based on the credit history in the bank and acquired from the Credit Bureau. Note that no collateral is involved in these loans. Under the credit guaranty, loans are classified into 2 major types based on the purpose. Consumption credit loans, under 500,000 RMB, are for individuals own consumption use, and business credit loans, under 1,000,000 RMB, are for small business purposes. Both types of loans are further classified into several categories.

Table 4 and Figure 2 show the details of loan categories. In table 4, life insurance, housing, and vehicles types are add-on loans to those borrowers who have had loans for these purposes previously. The amount and interest rate are decided based on the previous loan amount, time limit of the contract, and repayment records. Government official is a special type that uses the civil servants position as a guarantee of their repayment ability.

Before June 2015, loans given based on evaluating credit history and income for consumption type, or based on business revenue and tax payment for business type, are categorized as standard. After that, however, standard type is only used for business purpose loans. For consumption purpose, another category called income is introduced. The new income type is a revised and optimized version of the previous standard type. Based on the income measurement used in the standard type, a more refined credit rating model is included to measure a customers risk level.

In the data set, the length of maturity ranges from 3 months to 60 months. The annual interest rate ranges from 9.96% to 22.68%. The interest rate is not only decided by costumers potential risk of default but also it incorporates the potential attribution a costumer has to the bank. As a customer is more likely to create a tight bond with the bank, such as getting more loans in the future, the bank is able to lower the interest rate to build up a long-term relationship with the customer.

The interest rate is fixed over the lifetime of the loan, unless there are incidents that might impose real cost on both bank and household. Borrowers need to apply for adjustments of interest rate if they think they are going to be delinquent.

The repayment amount in each month is the same. The amount is based on the formula:

$$MonthlyRepayment = \frac{A\beta(1+\beta)^m}{(1+\beta)^m - 1} \tag{1}$$

where A denotes the total loan amount, β is the fixed monthly interest rate, and m is the time limit of the loan. The repayment consists of both interest payment and principal repayment. The interest payment can be calculated by multiplying the interest rate to the remaining principal amount, and the principal payment is calculated by subtracting interest from monthly repayment amount. Under this method, the proportion of interest in repayment decreases from month to month, while the proportion of principal increases.

3.3. Characteristics of Borrowers

Our main variable, emigration status, is inferred by comparing a borrowers birth place to the city where the loan was originated. Therefore, our definition of emigration treats everyone born outside the city as emigrants, regardless of how long they may have lived in that particular city. In some cases, a person may have changed his permanent residency to the city where he receives the loan, which implies his birth place may have little correlation with his default risk. However, without the information of hukou (permanent residency), birth place is the closest proxy to residency we have.

A borrower can exit the sample in two ways: pay off the loan or default. The dataset has the records of last past due date of the loan payment — the past due that was not paid back until the end of the sampled period. In other words, if a person pays back his loans in arrear, his late payment is not treated as default and therefore not observed in the data. Only loans with active past due records are observable and classified as default. Loans with no past due information are assumed to be in good standing. There is no contract starts prior to August 2013, but some contracts end after our sampled period. Note that if a contract ends Feb. 2017 and shows no past due record, it is treated as a censored observation because we do not have information if the borrower will default or pay off in future.

The characteristics information collected at the beginning of the contract includes birth place, age, sex, marital status, type of working company, education level, monthly income, home ownership, and years of employment in the current company. Most of the demographic factors have also been frequently studied in past research on default rate (Stepanova and Thomas 2001, Baesens et al. 2004 etc.). We also included some new variables. For instance, the type of institution where a borrower works can affect default rate, because some companies have lower turn-over rate than the others. Borrowers with mortgage loans are more likely to default than borrowers with owned houses. Years of employment is treated as an indicator of job security. As for the impact of income, as mentioned before, can be more complicated.

While marital status, working company, education level, income, and home ownership may change during the sampled period, we only observe them at the initiation of the loan. The summary statistics of some variables are shown in Table 5.

4. Methodology

To understand the implicit relationship between the probability of default and its relationship with immigration status, this study will employ both logistic regression and survival analysis. While logistic regression can help to determine the impact of migration status on the probability of default, survival analysis allows one to estimate its impact on the time to default. More detailed descriptions of the two methods can be found in the following sections.

Traditionally, a default is defined as the case of which every time a repayment of the lending institution is not met on time. Because of the limitation of the current dataset, however, only the last period of delinquent when the data is collected is recorded, and so previous delinquents are not considered in the model. With the record of the last delinquent of default loans, we are able to calculate the percentage of loans paid back.

4.1. Logistic Model

Logistic model is used here to understand the unconditional probability of default on borrowers. The logistic regression function is

$$P(x) = \frac{(e^{\beta X})}{1 + (e^{\beta X})} \tag{2}$$

where X is a matrix of all the independent variables, β is a matrix of the coefficients of the independent variables, and the value of P is between zero and one. To quantify the probability of default, a dummy variable is created to represent the event of whether a default does occur. When the dummy variable equals 0, default does not occur, and it is assigned with 1 when default does occur. The value of $P(x)$ falls between 0 and 1, representing the probability that a default will occur.

Since the logistic model is nonlinear, OLS is not applicable to estimate the unknown parameter. Maximum likelihood estimation (MLE) can be used to find the estimators of logistic model as it is a nonlinear binary response model. The MLE of β, denoted as $\hat{\beta}$, maximizes the log-likelihood, where equation 3 gives the log-likelihood for observation i. Equation 4 is the log-likelihood for a sample size of n, which is simply the sum of the value from the first equation for each observation. The $\hat{\beta}$ is thus the logit estimator for a logistic model.

$$l_i(\beta) = y_i \log\left[G(X_i\beta)\right] + (1 - y_i) \log\left[1 - G(X_i\beta)\right] \tag{3}$$

$$L(\beta) = \sum_{i=1}^{n} l_i(\beta) \tag{4}$$

Two models are set up to further study how migration status would affect a borrowers probability of default. Both Model 1 and Model 2 employ logistic regression that use default status as dependent variable, and other applicants characteristics as independent variables. To understand the effect of income level on probability of default of two groups of people, Model 2 differs from Model 1 by adding an interaction of migration status and monthly income.

The resulting table is going to have only the odd ratio, which is obtained by taking the exponential of the coefficients. The coefficients of the variables are regression coefficients of the linear predictor function that gives little information unless transformed into logistic regression, but they are the indicators of directions of change. To understand how default rate is being affected, the percentage change in the odds of default can be computed using (odds ratio -1) *100: the variable would increase the default rate if the result is positive, and decrease the default rate if negative.

4.2. Cox's Proportional Hazard (PH) Model

The aim of survival analysis is to estimate the probability density distribution of T:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T \leq t + \Delta t | T \geq t)}{\Delta t} \tag{5}$$

The Cox's proportional hazard model assumes

$$h(t) = e^{\beta X} h_0(t) \tag{6}$$

where $h_0(t)$ is the baseline hazard rate, and X denotes a vector of covariates. Different from parametric models, $h_0(t)$ can be any function of time, which is why Cox's PH model is a generalization of the parametric proportional hazard models. This feature gives the model an advantage over parametric models in terms of robustness and flexibility, because it is not necessary to specify which distribution is most appropriate for the baseline model. It is been shown that the coefficients can be estimated without any knowledge of the baseline hazard (Cox, 1972).

It is called proportional hazard model because at any point of time, the ratio of hazards of two individuals is a constant. Nevertheless, the model can also be easily extended to non-proportional hazard model. The estimation method proposed by Cox is called partial likelihood. This method only requires the knowledge of the order, not the exact time, in which the events occur. Partial likelihood estimators bear the merits of being consistent and normally distributed. Although they are not fully efficient, the loss of efficiency is often marginal (Efron, 1977).

5. Results

5.1. Logistic Model

The results of two logistic models,shown in the table 7, reveal how immigration status would affect the default rate with different income levels.

In model 1, monthly income and immigration status are treated as two isolated variables of which monthly income has a small negative coefficient indicating default rate would decrease as the income increases, while the probability of default would increase 2.39% if one is migrant, although the change is statistically insignificant. The result of model 1 without interaction is intuitive.

Model 2, however, provides more insight into the relationship of levels of income and immigration status by taking an interaction of two variables to examine the hypothesis proposed in the beginning. Immigration status and income levels are both shown statistically significant. With the interaction, the likelihood of default for non-migrants is solely influenced by income and other control variables, whereas immigration status and the interaction terms could also influence the likelihood of default for migrants. The coefficient for immigration status is positive but negative for income and interaction term, leading to a higher default rate for migrants when income is low, because the effect of the interaction term can not offset the effect of immigrant status. However, as income increases, the gap of default rate between migrants and non-migrants gradually decreases. Around 80th percentile of the income level, two groups exhibit the same probability of default. Beyond that critical point of income, the default rate is higher for non-migrants than for migrants.

Besides the interaction of immigration status and monthly income, other control variables are the same in both models.

5.2. Cox's Proportional Hazard Model

The results of Cox's PH model are shown in table 8. Model 1 and model 2 both use only the applicants characteristics as explanatory. Based on model 1, model 2 adds an interaction term of monthly income and migration status.

The hazard ratios reported are derived by exponentiating the estimated coefficients. While the coefficient itself tells little information, we can derive the percentage change in the hazard of default when explanatory increases one can by calculating (hazard ratio 1) x 100. Note that no intercept is reported because it is part of the undefined baseline model.

Model 1 tells us that the hazard of immigrants is 3.69% higher than that of non-immigrants, although the difference is insignificant at 0.1 significance level. Increase in monthly income has a small but significant reducing effect on the hazard of default. In other words, more income implies that the loan is more likely to survive a longer period of time. The result is intuitive since more income usually suggests less default risk.

In model 2, however, the immigration status and the interaction term are both significant. Although not showing in the table, the coefficient of the interaction term is slightly below zero. The results indicate that a low-income immigrant is likely to default earlier than a non-immigrant with the same income and the same other characteristics. However, the hazard of an immigrant will keep dropping as the income level increases. The income level of a non-immigrant, on the contrary, has little impact on the default hazard. Therefore, the income level may increase to a point where the hazard of immigrants is lower than the hazard of non-immigrants, with else things the same. In another word, if we compare to non-immigrants, poor immigrants have higher default risk, while rich immigrants have lower default risk. Indeed, with some simple calculation, our model shows that the intersection point happens at 45406.99 RMB of income (around 79th quantile): immigrants are more likely to default than non-immigrants when their income is below this level, but less likely to default when income is higher. Therefore, at median income level, immigrants are more likely to default. Model 2 also explains the reason why immigration status in model 1 is insignificant. While there are both poor and wealthy immigrants in the dataset, the immigration status in model 1 only shows the average result of the two groups.

Motivated by the research of Wang et al.(2015), here we provide one possible explanation of the relationship between migration status and income level. Immigrants in the lower income tail usually cluster at labor-demanding jobs. Without local *hukou*, these people are not secured with stable job positions and can easily be substituted. Moreover, theyre suffered from more pressure by sending back remittance to their family. Immigrants in the upper income distribution, however, are positively selected by the job market. These people are usually skilled workers and are hard to be substitute. In addition, they have strong incentive to work harder than non-immigrants so they will not be discriminated in the job market.

Figure 4 shows the obtained baseline model of Model 2: the survival probability as time goes by. As we can see, the baseline hazard is also a step function, with a particular considerable drop at around 38th month.

6. Conclusion

Our study employed a dataset from a commercial bank in China to explore the impact of migration status on default risks on credit loans. Although loan default is a well-developed research field, our topic is particular interesting because the role of migration has not been studied before. Most studies on migration agree that migrants are faced with discrimination in terms of both job-searching and social welfare. The considerable financial burden implies that migrants are more likely to default than local people. However, the issue becomes more complicated when we consider income level. The cost of migration itself selects out 2 groups of migrants: one group with great skills and endowment to survive in the city, and another group who are willing to do low-tier jobs shunned by local people. The first group, usually with high income, are favored by employers over local workers. The second group work at jobs characterized by high mobility and low job security. Therefore, we hypothesize that migrants at low income level are more likely to default than non-immigrants. As the income level increases, the default risk of migrants decreases, and may even be lower than local people at some point.

Using both logistic and Cox proportional hazard model, we studied both the default probability and the time to default, and find that the results are consistent to our hypothesis. Specifically, when income level reaches around 80th percentile, immigrants and non-immigrants have about the same default risk.

This study shows that migration status indeed impacts the default risk. However, we recognize that the results can be inaccurate, since our sample is collected from borrowers of a bank instead of the entire population.

7. References

Baesens, B., Van Gestel, T., Stepanova, M., Van den Poel, D., & Vanthienen, J. (2005). Neural network survival analysis for personal loan data.Journal of the Operational Research Society,56(9), 1089-1098.

Banasik, J., Crook, J. N., & Thomas, L. C. (1999). Not if but when will borrowers default.Journal of the Operational Research Society, 1185-1190.

Cox, D. R. Regression Models and Life-Tables.

Efron, B. (1977). The efficiency of Cox's likelihood function for censored data.Journal of the American statistical Association,72(359), 557-565.

Fan, C. C. (2003). Ruralurban migration and gender division of labor in transitional China.International Journal of Urban and Regional Research,27(1), 24-47.

Fan, C. C. (2007).China on the Move: Migration, the State, and the Household. Routledge.

Ma, L. J. (2002). Urban transformation in China, 19492000: a review and research agenda.Environment and planning A,34(9), 1545-1569.

Meng, X., & Zhang, J. (2001). The two-tier labor market in urban China: occupational segregation and wage differentials between urban residents and rural migrants in Shanghai.Journal of comparative Economics,29(3), 485-504.

Shi, L. (2008).Rural migrant workers in China: scenario, challenges and public policy. Geneva: ILO.

Song, Y., Zenou, Y., & Ding, C. (2008). Let's not throw the baby out with the bath water: the role of urban villages in housing rural migrants in China.Urban Studies,45(2), 313-330.

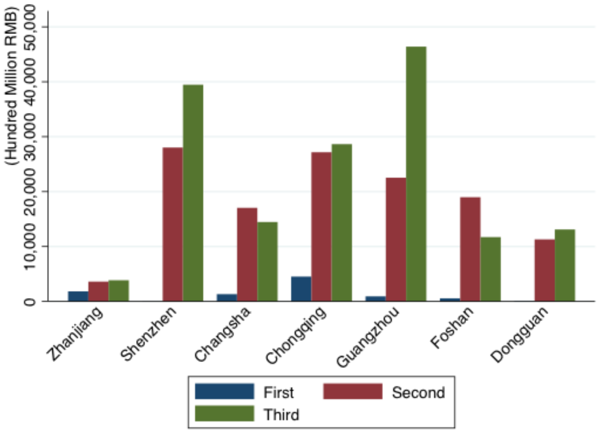
Stepanova, M., & Thomas, L. C. (2001). PHAB scores: proportional hazards analysis behavioural scores.Journal of the Operational Research Society, 1007-1016.

Sun, M., & Fan, C. C. (2011). China's permanent and temporary migrants: differentials and changes, 19902000.The Professional Geographer,63(1), 92-112.

Wang, H., Guo, F., & Cheng, Z. (2015). A distributional analysis of wage discrimination against migrant workers in Chinas urban labour market.Urban Studies,52(13), 2383-2403.

8. Appendix

Figure 1: Average GDP Generated from the 3 Categories of Industry during 2013-2016 in Cities where the Bank Locates



Note: Industries in China are typically classified into 3 major categories in China. The first category includes agriculture, forestry, livestock production, and fishing. The second category includes mining, manufacturing, construction, production of electric power, heat, gas, and water. The third category includes industries other than the first two categories, and it is mainly characterized by service jobs and technological development.

Figure 2: Distribution of loans

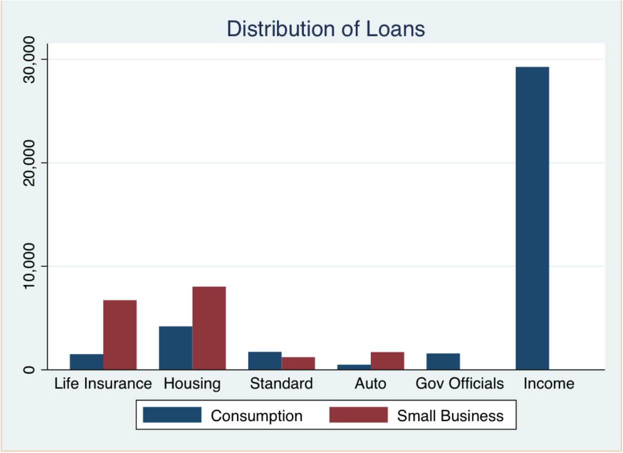


Figure 3: Monthly Income Distribution Comparison between Non-immigrants and Immigrants (excluding borrowers with top 5% monthly income)

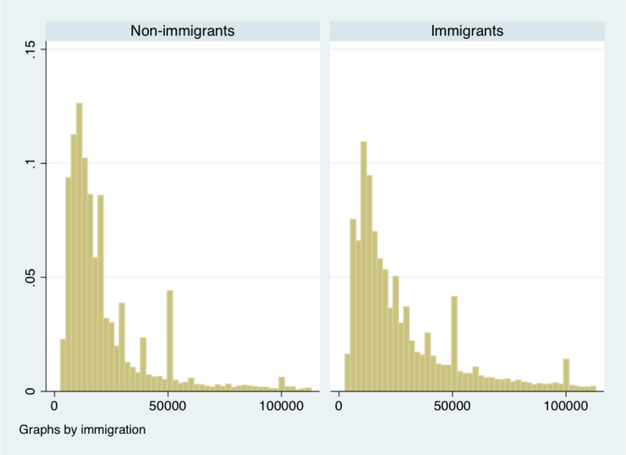


Figure 4: Baseline Model of Cox's Proportional Hazard Model

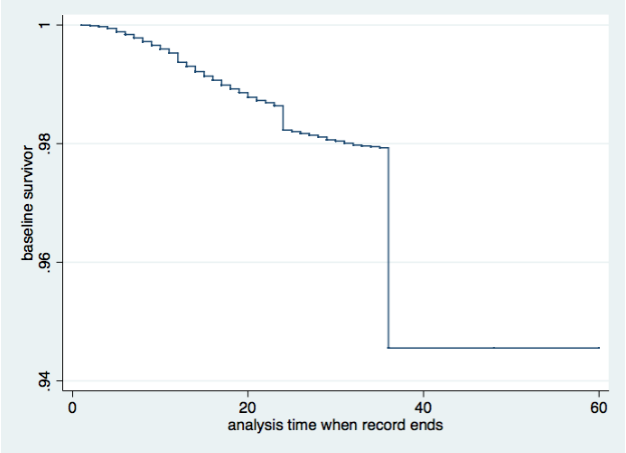


Table 1: Characteristics of Cities Where the Bank Has Branches.

City	Land Area (km2)	Populationx (million)	Real Disposable Income per capita (RMB)
Zhanjiang	11,693	7.22	15883.10
Guangzhou	7,434	13.39	32912.42
Shenzhen	1,997	11.17	44732.32
Changsha	11,819	7.40	34247.06
Chongqing	82,400	30.07	20025.24
Foshan	3,875	7.38	36881.07
Dongguan	2,465	8.29	40714.75

Data Source: *National Bureau of Statistics of China*

Notes: Both population and real disposable income show the average amount during year 2013 to 2016. Population of any year equals to the number of people who have lived in the city for over 6 months in that particular year, regardless their migration status. Real disposable income is inflated to year 2013.

Table 2: The Time of Establishment and The Number of Sub-branches in 7 Cities

City	Number of Sub-branches	Year of the First Sub-branch Established
Zhanjiang	61	1998
Guangzhou	10	2009
Shenzhen	10	2010
Changsha	9	2010
Chongqing	9	2010
Foshan	8	2011
Dongguan	6	2011

Data Source: *Website of the Commercial Bank*

Table 3: Descriptive Statistics of Loans

Variable	Min	Max	Mean	Std.
Starting Month	Aug. 2013	Feb. 2017		
Time Limit of the Contract (months)	3	60	37.7760	11.9781
Loan Amount (RMB)	10,000	500,000	164482.8	121077.8
Annual Interest Rate (%)	0.18	22.68	18.6742	3.52
Default Status[1]	0	1	0.0489	0.2156
Past Due (days)[2]	1	1207	141.481	175.4922

Data Source: *Dataset Provided by the Commercial Bank*

[1] For default status, 0 means no default record observed, and 1 means the loan that is at least one day past due. The mean value suggests that 4.89% (2651 out of 56,269) of all loans have past due records. Out of the all loans with default records, 901 loans are less than 30 days past due, 304 loans are 30-60 days past due, and 1,446 are greater than or equal to 60 days past due.

[2] Exclude loans with no past due.

Table 4: Types of Loans

Category	Consumption	Small Business
Life Insurance	✓	✓
Housing	✓	✓
Standard[1]	✓	✓
Vehicles	✓	✓
Government Officials[2]	✓	
Income[3]	✓	

Data Source: *Dataset Provided by the Commercial Bank*

Notes: a person can borrow credit loan for either consumption purpose or for his own small business. Credit loans for small business are distinct from loans for companies or enterprises, since it is guaranteed by personal credit.

[1] (need more info)

[2] Government officials have the privilege to borrow with less restrictions on guarantee.

[3] The bank assess the borrower's payment ability based on his salary or wage.

Table 5: Characteristics of Borrowers

	Min	Max	Mean	Std.
Immigration[1]	0	1	0.7597	0.4272
Age	22	56	36.7888	7.6424
Sex[2]	0	1	0.6969	0.4596
Education Level[3]	0	4	1.6869	1.2334
Monthly Income (RMB)	2,700	8,925,000	36705.98	80368
Monthly Income Excluding Top 5%	2,700	113,064	26588.7	22401.15

Data Source: *Dataset Provided by the Commercial Bank*

Notes: The top 5% income deviates substantially from the average level. We suspect the outliers are caused by recording errors.

[1] 0=non-immigrant; 1=immigrant.

[2] 0=female; 1=male.

[3] 0=middle school; 1=high school or technical secondary school; 2=junior college; 3=college; 4=graduate school.

Table 6: Summary Statistics of Immigrants and Non-immigrants

	Mean of Non-Immigrants	Mean of Immigrants
Number of Observations	13,039	41,230
Default Status*	0.0409	0.0514
Past Due (days)*	4.7445	7.5965
Time to Default (Months)*	39.9709	35.7490
Loan Amount (RMB)*	151527.7	168579.8
Term of Contract (Months)*	40.8940	36.7899
Interest Rate (%)*	17.8143	18.9462
Share of Female*	0.6596	0.7087
Age*	37.4684	36.5738
Income (RMB)*	28921.099	39167.954
Income excluding top 5% (RMB)*	22192.39	28014.83
Years in Current Company		
Marital Status		
Single*	0.1564	0.1698
Married	0.7692	0.7645
Divorced*	0.0719	0.0638
Others	0.0025	0.0020
Occupation		
Personal Business*	0.0726	0.1398
Company*	0.4799	0.1292
Others*	0.0171	0.0319
Army/Police	0.0009	0.0004
State-owned Enterprises*	0.1618	0.0871
Private-owned Enterprises*	0.1491	0.3902
Joint-Stock Enterprises*	0.1186	0.2213
Housing Type		
Rental Home*	0.0278	0.3569
Mortgage*	0.3657	0.3497
Self-Owned*	0.2742	0.1122
Joint-Owned	0.0005	0.0004
Dorm*	0.0133	0.1094
Relative*	0.1014	0.0304
Others*	0.2158	0.0415
Loan Type: Consumption		
Gov Officials*	0.0425	0.0232
Life Insurance*	0.0147	0.0312
Housing*	0.0650	0.0768
Standard	0.0503	0.0243
Income*	0.6614	0.4724
Auto	0.0090	0.0087
Loan Type: Business		
Life Insurance*	0.0295	0.1485
Housing*	0.0877	0.1607

Standard*	0.0137	0.0244
Auto*	0.0269	0.0321
Education		
Middle School	0.1873	0.3317
High school or technical secondary school	0.0214	0.0562
Junior college	0.3899	0.3424
College	0.3820	0.2397
Graduate school	0.0194	0.0299
Location of the Bank		
Dongguan*	0.0632	0.2738
Foshan*	0.0033	0.0021
Guangzhou*	0.2344	0.1809
Shenzhen*	0.0718	0.3702
Zhanjiang*	0.4782	0.0156
Chongqing*	0.1164	0.1480
Changsha*	0.0327	0.0094

***** The mean values between non-immigrant and immigrant groups are different at 0.05 level.

Data Source: *Dataset Provided by the Commercial Bank*

Notes: The mean value of each categorical variable denotes the percentage of (non-)immigrants in that category of all (non-)immigrants.

Table 7: Logistic regression using only applicants' characteristics variables

	Model 1	Model 2
	Odds Ratio	Odds Ratio
Dependent Variable		
Default status	.	.
Independent Variables		
Immigration status	1.0239	1.1760**
Monthly income	1.0000***	1.0000**
Immigration x Monthly income		1.0000***
Loan amount (RMB)	1.0000***	1.0000***
Time limit of loan	1.0087***	1.0088***
Annual interest rate (%)	1.1445***	1.1450***
Sex	0.9984	0.9977
Age of borrower	1.0021	1.0019
Years worked in current job	1.0049*	1.0047*
Location of Bank		
Foshan	0.3979	0.3981
Guangzhou	2.4395***	2.4282***
Shenzhen	1.3245***	1.3225***
Zhanjiang	1.8741***	1.9385***
Chongqing	3.3860***	3.3984***
Changsha	0.6853	0.6898
Marital Status		

married	0.8441***	0.8446***
divorced	1.0858	1.0840
others	0.4934**	0.4940**
Occupation Type		
company	0.9234	0.9328
others	0.6264***	0.6287***
army/police	0.5846	0.5852
state-owned enterprise	0.8704*	0.8765*
private enterprise	1.0893**	1.0917**
joint-stock enterprise	1.0300	1.0318
Education Level		
high school or technical secondary school	1.1558**	1.1564**
junior college	0.8869***	0.8861***
college	0.8052***	0.8028***
grad school	0.6539***	0.6503***
Housing of the Borrower		
mortgage	1.1031**	1.1128**
owned home	1.0376	1.0472
jointly owned home	0.9690	0.9946
dorm	1.1101	1.1057
relative's home	0.9608	0.9777
others	1.2140***	1.2309***
Loan Type: Consumption		
life insurance	1.2508**	1.2644**
housing	1.3761***	1.3865***
standard	5.0464***	5.1368***
income	0.2126***	0.2162***
auto	2.0056***	2.0186***
Loan Type: Business		
life insurance	1.2692**	1.2849**
housing	1.6717***	1.6863***
standard	0.4460***	0.4479***
auto	1.5697***	1.5820***
Constant	0.0067***	0.0059***
Observations	51,446	51,446

******* p<0.01, ****** p<0.05, ***** p<0.1
Data Source: *Dataset Provided by the Commercial Bank*
Notes: Both model 1 and model 2 use only the information of the applicants. Model 1 does a simple logistic regression, where default status is the dependent variable and applicants' characteristics variables are independent variables. Model 2 add an interaction term of immigration status and monthly income, allowing the effect of immigration status to vary across income levels. The percentage change in the odds of default can be computed using (odds ratio – 1) x 100. The model excludes the 5% borrowers with the highest income and another 90 observations with recording errors or missing value.

Table 8: Cox's proportional hazard model using only applicants' characteristics variables

	Model 1 Hazard Ratio	Model 2 Hazard Ratio
Dependent Variable Time to Default	.	.
Independent Variables		
Immigration status	1.0369	1.1559**
Monthly income	1.0000***	1.0000
Immigration x Monthly income		1.0000**
Loan amount (RMB)	1.0000	1.0000
Annual interest rate (%)	1.0468***	1.0469***
Sex	0.9882	0.9879
Age of borrower	1.0001	0.9999
Years worked in current job	1.0039	1.0036
Location of Bank		
Foshan	0.3172	0.3172
Guangzhou	1.8674***	1.8596***
Shenzhen	1.2190***	1.2174***
Zhanjiang	1.7555***	1.8036***
Chongqing	2.3662***	2.3713***
Changsha	0.5728**	0.5751**
Marital Status		
married	0.8910***	0.8912***
divorced	1.0720	1.0704
others	0.5727*	0.5746*
Occupation Type		
company	0.9281	0.9350
others	0.8076**	0.8097*
army/police	0.6953	0.7026
state-owned enterprise	0.8885*	0.8928*
private enterprise	1.0920**	1.0934**
joint-stock enterprise	1.0113	1.0125
Education Level		

high school or technical secondary school	1.1304**	1.1304**
junior college	0.9061***	0.9053***
college	0.8370***	0.8349***
grad school	0.7211***	0.7181***
Housing of the Borrower		
mortgage	1.0720*	1.0787*
owned home	1.0608	1.0681
jointly owned home	0.8915	0.9266
dorm	1.0595	1.0560
relative's home	0.9992	1.0123
others	1.1750***	1.1860***
Loan Type: Consumption		
life insurance	1.2864**	1.2965**
housing	1.3402***	1.3482***
standard	3.0469***	3.0863***
income	0.2359***	0.2388***
auto	1.9253***	1.9357***
Loan Type: Business		
life insurance	1.6770***	1.6922***
housing	2.0535***	2.0669***
standard	0.6860**	0.6906**
auto	1.9078***	1.9207***
Observations	51,446	51,446

*** p<0.01, ** p<0.05, * p<0.1

Data Source: *Dataset Provided by the Commercial Bank*

Notes: Both model 1 and model 2 use only the information of the applicants. The dependent variable is the time to default. The independent variables include only the applicants' characteristics. Model 2 adds an interaction term between monthly income and migration status to model 1, allowing the effect of immigration status to vary across income levels. The hazard ratios are the exponentiated coefficients. No intercept is reported because it is part of the undefined baseline model. The percentage change in the hazard of default when explanatory increases one can be computed by (hazard ratio – 1) x 100. The model excludes the 5% borrowers with the highest income and another 90 observations with recording errors or missing value.