Minh Vu

November 04, 2020

Econ Honors

**Writing Assignment #2**

Research Question: What are the determinants of the microcredit loan default rate?

# Introduction

I have always been fascinated with data and its application in our modern world. Data is the new fuel of the 21st century; it has improved a wide range of industries, such as medical, computer science, finance, etc. As a student double majoring in Computer Science and Economics and minoring in Data Science, I am curious to know more about the application of data, or Machine Learning to be more exact, in those two fields. I want to know more about loans and to gain more insight into different factors that might affect the default rate in the retail banking industry and non-bank financial institutions.

Credit scoring is a key risk assessment technique to analyze and quantify a potential obligor’s credit risk. Essentially, credit scoring aims at quantifying the likelihood that an obligor will repay the debt. The outcome of the credit scoring exercise is a score reflecting the creditworthiness of the obligor. In the last couple of decades, Machine Learning has been playing a crucial role in building Credit Risk models, or also generalized as Scorecard models. The input to the Scorecard model in credit risk includes information contained in the client's profile. They are collected mainly from the credit bureau or data warehouse of the bank. The data is divided into these four basic groups:

* Demographic: Information related to personal characteristics such as education level, income, sex, age, occupation, marital status, family size, number of dependents, etc.
* Credit history: Information that is managed at the credit bureau. Customer loan history data from all banks operating in the territory of a country is aggregated into one data center. Thus, the bank can cross check customers' credit information from other banks.
* Transaction information: Transaction history on credit cards or ATM cards will partly evaluate the financial capacity of the customer. Hence this information is very useful for forecasting solvency.
* Collateral information: Information that comes with the mortgage loan. The value of collateral will be the means to cover the risk in case the customer defaults.

There has been a lot of research about determinants of loan default in the finance industry. However, not all the topics of research are the same and it is important to understand the difference between them. There are three parts of this topic that can create many variations: the type of determinants, the type of loan and the type of organization.

There are two main types of determinants: individual’s characteristics and macro determinants. A wide range of macro variables such as GDP per capita, GPA growth rate, unemployment rate, etc., have been studied for years to understand their impact on non-performing loans. However, I am not looking into these factors. In the span of this research, I only focus in individual’s characteristics, consisting of demographic data such as sex, age, income, occupation, etc., and previous loan repayment information. With regard to the type of loan, the majority of the observations in the chosen dataset are small loans, belonging to one of these three categories: revolving loan (credit card loan), point of sales loan and cash loan. Finally, the finance industry comprises many different types of organizations such as retail banks, commercial bank, microfinance institution, non-bank financial institution, etc. The chosen dataset is from Home Credit – a non-bank financial institution.

According to Balogun and Alimi (1988), loan default can be defined as the inability of a borrower to fulfil his or her loan obligation as at when due. Ntiamoah, Oteng, Batrice Opoku and Siaw (2014) describe the cause of this problem as “multi-dimensional” and “non-uniform among different literatures”. Microfinance perceived as a financially sustainable instrument meant to reach a significant number of poor people of which most are not able to access financial service because of the lack of financial records, limited credit history and lack of assets for collateral. As a result, this is a highly risky business since it involves high screening, monitoring and enforcement costs.

Some of the loans given out by the lending institutions, unfortunately, are not paid back and eventually result in bad debts with adverse consequences for the overall financial performance of the institutions. The costs of loan delinquencies would be felt by both the lenders and the borrowers. The lender has costs in delinquency situations, including lost interest, opportunity cost of principal, legal fees and related costs. For the borrower, the decision to default is a trade-off between the penalties in lost reputation from default versus the opportunity cost of forgoing investments due to working out the current loan.

Interestingly, while the poor has been generally perceived to have low credit worthiness, their repayment rates of the loan has been generally quite impressive. It has been reported that the loan repayment rates of Grameen Bank in Bangladesh are almost always above 95% (Morduch, 1999). In Malaysia, the repayment rates of Amanah Ikhtiar Malaysia (AIM), which is a modified replication of the Grameen Bank, is about 97% (BNM, 2006). The low default rates of these microcredit programs have led observers to believe that lending to the poor and the low income group of microenterpreneurs might not be as risky as it has been conventionally presumed (Roslan and karim, 2014).

# Literature Review

## The impact of demographic factors on the default rate

Roslan and Karim (2009) investigated the effect of borrowers characteristics, projects characteristics and loan characteristics on loan repayment of Agrobank Micro Credit Scheme - a commercial bank in Malaysia. The topic of research is closely related to my research question. The Agrobank Microcredit Scheme is also a non-group lending based with flexible repayment schedule and longer repayment period. In this study, the characteristics of borrowers are (i) gender; (ii) marital status; (iii) race; (iv) level of education; (v) age; (vi) occupation; (vii) number of dependents; (viii) experience; (ix) membership in business society and (x) training. Roslan and Karim (2009) found that, among all demographic variables, the probability for default is mostly influenced by the gender of the borrower, type of business activity, and training. The probability to default is found to be lower among female borrowers as compared to male borrowers. With regard to the type of industry the borrower works in, they discovered that borrowers involved in services/support activities have a lower probability to default compared to those in production activities. In addition, the analysis shows that borrowers who did not have any training (in relation to their business/project activity) have a higher probability to default compared to those borrowers who had some training.

Another study was conducted in Malaysia by Mokhtar, Nartea and Gan to investigate the determinants of loan repayment problems. However, there is a slight difference in the type of organization in their study compared to Roslan and Karim’s: they looked into loans from *microfinance* institutions - TEKUN and YUM. Yet, they also found that male borrowers in TEKUN had problems in repaying their loan. With respect to the type of industry both TEKUN and YUM borrowers work in, involving the agricultural businesses contributed to loan repayment problems. In addition, the study found that the age of the borrower is also a crucial factor that worsen the loan repayment situation. TEKUN and YUM borrowers aged between 46 to 55 years old and 18 to 25 years old, respectively, had loan repayment problems.

Some of the same discoveries on demographic traits were also made by Setargie (2013) in his research in Ethiopia. He made some similar observations. Setargie (2013) pointed out that the most significant determinants of credit default performance were education, credit/loan size, credit diversion, availability of other credit sources, credit/loan supervision, suitability of credit repayment period and income. Like Roslan and Karim, Setargie also found that being male reduces the credit repayment performance. Education and income are found to decrease the probability of credit default.

Just like Roslan and Karim (2009), Wongnaa and Awunyo-Vitor (2013) also looked into loans from *financial institutions* in their study in Ghana. They learned that educational level, number of years of farming experience, profit gained from loan, age of farmer, supervisory visits to farmers and access to off-farm income have positive effects on yam farmers’ ability to repay the loans given to them by financial institutions. A rise in each of these factors will therefore enhance yam farmers’ loan repayment abilities. On the other hand, being male and being married have negative effects on yam farmers’ loan repayment abilities. In another study conducted on farmers, Akinwemi and Ajayi (1990) fount out that farm size, family size, scale of operation, family living expenses and exposure to sound management techniques were some of the factors that can influence the repayment capacity of farmers.

## The impact of loans types on the default rate

In a study of determinants of rural loan repayment performance in Negeria, Olomola (1999) discovered that loan disbursement lag and high interest rate can significantly increase borrowing transaction cost and, therefore, adversely affect repayment performance. Agreeing with Olomola, Okpugie (2009), who was also looking into data from microfinance banks in Negeria, and Vandel (1993), who was looking into data of mortgage loan, both confirmed that high interest rate charged by the microfinance banks has been discovered to be the main reason behind the alarming default.

Other loan characteristics, such as the size of the loan and the loan repayment period, are also found to be important to avoid loan repayment deliquency. Based on a study of loan repayment of a commercial bank in Malaysia, Roslan and Karim (2009) suggested that the probability of default is higher as the smaller the size of the loan and the longer the repayment period. Another study also based in Malaysia by Mokhtar, Nartea and Gan shared the same discoveries. While weekly loan repayments was found to cause some problems, loan repayment period over one year improved the probability of default.

Setargie (2013), aside from the same discovery about loan size as above, observed that credit diversion was one of the significant factors influencing credit default negatively. Although this is an individual’s characteristic, it is heavily influenced by many loan’s characteristics, of which loan supervision and suitability of repayment period were found to reduce the probability of diverting credit to non-productive uses that ultimately lead to reduced recovery rate.

# Data

## Home Credit Default Risk: Can you predict how capable each applicant is of repaying a loan?

  This dataset is provided by Home Credit – a non-bank financial institution specializing in providing loans to people that have low exposure to banking services in four main geographical areas: Central and Eastern Europe (CEE), Commonwealth of Independent States (CIS), China and South & South East Asia (SSEA). This fact leads to some interesting points worth noticing before conduting a study on this dataset:

* Home Credit mainly does business in CIS and SEA countries so the data might be combined from data of Kaz, Russia, Vietnam, China, Indonesia, Phillipines. Therefore, FICO score from US is really not useful.
* Bureau data is usually not sufficient for credit scoring loan application here, as people have low exposure to banking.
* Current loan & Previous loan are more reliable data for scoring

The dataset comprises of 7 smallers tables, connecting with each other in this relationship



Figure 1: Data Map

* ***application\_{train|test}.csv***
  + This is the main table, broken into two files for Train (with TARGET) and Test (without TARGET).
  + Static data for all applications. One row represents one loan in our data sample.
  + Train table has more than 300,000 observations, each with 122 features
* ***bureau.csv***
  + All client's previous credits provided by other financial institutions that were reported to Credit Bureau (for clients who have a loan in our sample).
  + For every loan in our sample, there are as many rows as number of credits the client had in Credit Bureau before the application date.
* ***bureau\_balance.csv***
  + Monthly balances of previous credits in Credit Bureau.
  + This table has one row for each month of history of every previous credit reported to Credit Bureau – i.e the table has (#loans in sample \* # of relative previous credits \* # of months where we have some history observable for the previous credits) rows.
* ***POS\_CASH\_balance.csv***
  + Monthly balance snapshots of previous POS (point of sales) and cash loans that the applicant had with Home Credit.
  + This table has one row for each month of history of every previous credit in Home Credit (consumer credit and cash loans) related to loans in our sample – i.e. the table has (#loans in sample \* # of relative previous credits \* # of months in which we have some history observable for the previous credits) rows.
* ***credit\_card\_balance.csv***
  + Monthly balance snapshots of previous credit cards that the applicant has with Home Credit.
  + This table has one row for each month of history of every previous credit in Home Credit (consumer credit and cash loans) related to loans in our sample – i.e. the table has (#loans in sample \* # of relative previous credit cards \* # of months where we have some history observable for the previous credit card) rows.
* ***previous\_application.csv***
  + All previous applications for Home Credit loans of clients who have loans in our sample.
  + There is one row for each previous application related to loans in our data sample.
* ***installments\_payments.csv***
  + Repayment history for the previously disbursed credits in Home Credit related to the loans in our sample.
  + There is a) one row for every payment that was made plus b) one row each for missed payment.
  + One row is equivalent to one payment of one installment OR one installment corresponding to one payment of one previous Home Credit credit related to loans in our sample.

Since I am doing this thesis both for Economics and Computer Science majors, I am building two separate models, one for each major. For Economics, for the sake of simplicity of the study, I am only going to use the main applicaion\_{train|test} table to build Logistic Regression and Linear Regression models. For Computer Science, I am going to join the main applicaion\_{train|test} with the three repayment tables (POS\_CAH\_balance.csv, instalments\_payments.csv, and credit\_card\_balance.csv) to build several machine learning models. I am not going to use bureau.csv and bereau\_balance.csv in this study. There are several reasons for why I am proceeding with this plan.

1. As I have mentioned above, bureau data is usually not sufficient for credit scoring loan application here as people have low exposure to banking. As a result, bureau.csv and bereau\_balance.csv will be discarded.
2. Economics major prioritizes deep understanding of the topic over sophisticated, high-accuracy models. Therefore, Logistic Regression and Linear Regression models are good options because although they do not usually yield an accuracy as high as other more complexed machine learning models, they have a great interpretability, which help us understand the contribution of each independent variable to the dependent variable (TARGET).
3. Computer Science will prioritize deep unstanding of the technical parts: cleaning/tidying data, dimension reduction, feature engineering, and building machine learning models. Although those models may not have high interpretability, they will yield higher accuracy.

In the scope of the Economics study, I will only talk about applicaion\_{train|test} table.

## applicaion\_{train|test} table

This is the main table, consisting of information of more than 300,000 loan applicants. There are 122 columns for the Train dataset, including a TARGET column with two values of 1 – the applicant had some problem repaying the loan and 0 – the applicant had no problem repaying the loan. Some most important features, in my opinion, are type of loan (NAME\_CONTRACT\_ TYPE), loan size (AMT\_CREDIT), loan annuity (AMT\_ANNUITY), income (AMT\_INCOME\_ TOTAL), gender (CODE\_GENDER), occupation (OCCUPATION\_TYPE), age (DAYS\_ BIRTH), years employed (DAYS\_EMPLOYED), etc.

There are three main types of loan in the application\_train table: revolving loan, consumer installment loan and installment cash loan.

* Revolving loan (credit card):
  + Loan applicant is given a credit limit, he/she can spend/withdraw in a month within that credit limit, and at the end of the month Home Credit will inform him minimum payment he needs to make.
  + This is very popular to US and Western Europe countries as banking services developed, but not really much in Home Credit market. As a result, there are very few instances of this type of loan for both application and previous\_application dataset.
* Consumer installment loan (Point of sales loan – POS loan):
  + Loan applicant is given a credit limit to buy a goods (phone, laptop). For example, if the price of the goods (AMT\_GOODS\_PRICE) is 100 USD, he will need to pay out of pocket (AMT\_DOWN\_PAYMENT) 20 USD. 80 USD will be the loan from Home Credit that paid directly to goods seller. Loan applicant takes home the goods.
  + Loan applicant will need to repay that credit monthly, 30 days interval each month. The date installment needs to be paid each month is called DUE\_DATE. Monthly repayment (AMT\_INSTALMENT) includes principle and interest, which is calculated to make the repayment equally. The formula is the same with mortgage loan. Monthly repayment is also called AMT\_ANNUITY in the application dataset while whole principle of the loan is called AMT\_CREDIT.
  + This is very popular in Home Credit market since the loan is small, short, purposeful and very easy to get.
* Installment cash loan:
  + Loan applicant is given a lump sum of cash. He/she can spend for whatever purpose.
  + Loan applicant will need to repay that credit monthly, 30 days interval each month. The date installment needs to be paid each month is called DUE\_DATE. Monthly repayment (AMT\_INSTALMENT) includes principle and interest, which is calculated to make the repayment equally. The formula is the same with mortgage loan. Monthly repayment is also called AMT\_ANNUITY in the application dataset while whole principle of the loan is called AMT\_CREDIT.
  + This is very popular in Home Credit market since cash is king of payment method here. However, it is riskier than credit loan.

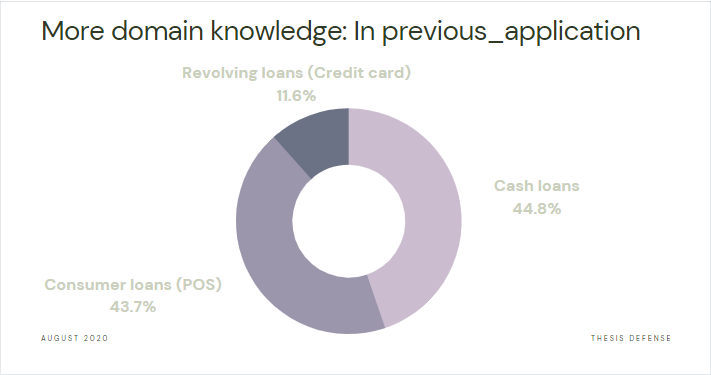
**

Figure 2: Types of loan

The loan size is reported in Russian Ruble and it ranges from 45,000 (~590 USD) to more than 4,000,000 (~52,000 USD). The loan size distribution has the mean of 599,026 (~7,800 USD), the median of 513,531 (6,700 USD), and the 75 percent quantile is 808,650 (~10,577 USD). The outlier threshold (Q3 + 1.5\*IQR) is 1,616,625 (~21,145 USD) and there are 6,562 instances above that threshold (~2% of the total dataset).

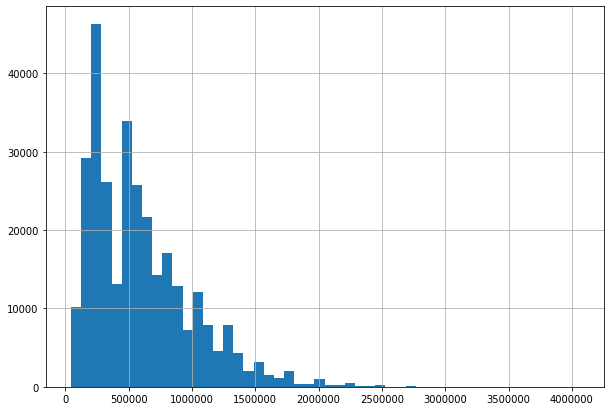


Figure 3: Loan size distribution before dealing with outliers

Since there is around two percent of the observations with loan size above the threshold, the threshold needs to be higher. The threshold for outlier is now 99.9 percent quantile, which is 2,517,300 (~ 32,926 USD). After dealing with outliers by capping the loan size at the new threshold, the distribution of loan size looks like this:

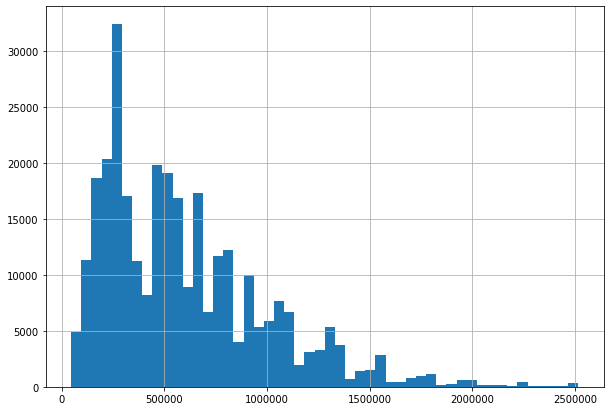


Figure 4: Loan size distribution after dealing with outliers

Income of the applicant is also a crucial factor in deciding the default probability. The income is reported in Russian Ruble and it ranges from 25,650 (~335 USD) to more than 117,000,000 (~1,530,360 USD). The income distribution has the mean of 168,797 (~2,207 USD), the median of 147,150 (1,924 USD), and the 75 percent quantile is 202,500 (~2,648 USD). The outlier threshold (Q3 + 1.5\*IQR) is 337,500 (~4,414 USD) and there are 14,035 instances above that threshold (~4% of the total dataset). Similar to loan size, the threshold of income needs to be higher. The threshold for outlier is now 99.9 percent quantile, which is 900,000 (~ 11,772 USD). After dealing with outliers by capping the income at the new threshold, the distribution of income looks like this:

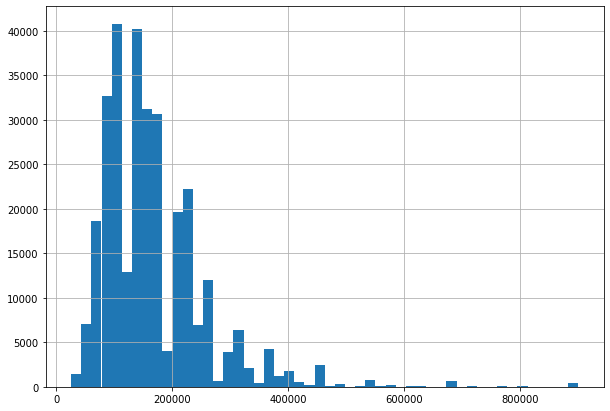


Figure 5: Income distribution after dealing with outliers

With respect to the probability of default, most applicants had no trouble repaying their loan.

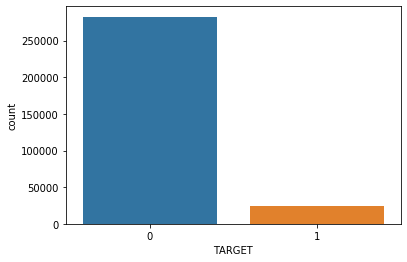


Figure 6: Default count

There are only two types of loans in the loan type feature of application\_train dataset since it merges point of sales loan and installment cash loan together as Cash loans. The pie charts below shows the percentage of default cash loan (cash\_1), default revolving loan (revolving\_1), paid back cash loan (cash\_0) and paid back revolving loan (revolving\_1).

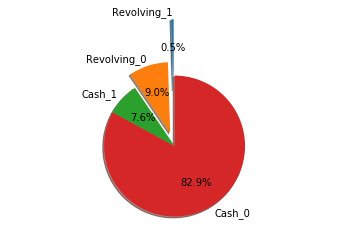


Figure 7: Default percentage of each type of loan

The plot below shows the relation between the size of the loan, the income of the applicant and the default rate. As can be seen, most loans that have some troubles in being repaid are small loans. It confirms an observation made by many reviewed papers above, which is the size of the loan has a negative relation with the default probability.

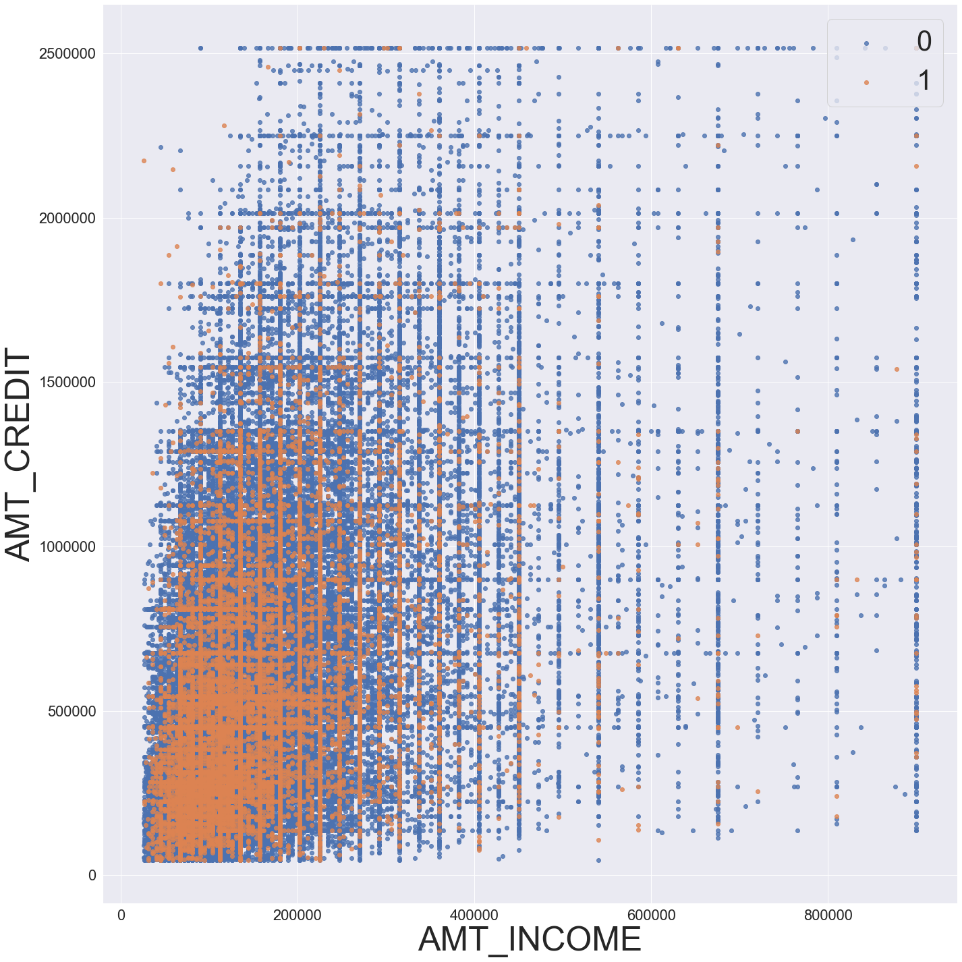


Figure 8: The relation between loan size, income and target

# Method

I will be testing the hypothesis:

H0: Default risk is not affected by any factor

H1: Default risk is affected by some factors

In order to prove that H0 is not significant, I will be using multilinear regression analysis. The base model is in the form below. The exact independent variables are still undecided; however, the majority of them will be important features mentioned above.

TARGET = β0 + β1\*income + β2\*loan size + β3\*loan annuity + β4\*education + β5\*occupation + β6\*gender + β7\*age + etc. + u

As I have mentioned above, although the multilinear regression model does not yeild as high of an accuracy as other Machine Learning models, it has great interpretability, which will offer us a better understanding of how each factor affects the default risk. After that, I will build a simple binary classifier using Logistic Regression to estimate the probability of each instance belonging to a class, with an equation as below:

Sigmoid Function

as z: the output of the multilinear equation

Table of Contents

[Introduction 1](#_Toc55919439)

[Literature Review 2](#_Toc55919440)

[The impact of demographic factors on the default rate 2](#_Toc55919441)

[The impact of loans types on the default rate 3](#_Toc55919442)

[Data 4](#_Toc55919443)

[Home Credit Default Risk: Can you predict how capable each applicant is of repaying a loan? 4](#_Toc55919444)

[applicaion\_{train|test} table 6](#_Toc55919445)

[Method 11](#_Toc55919446)

[Figures 12](#_Toc55919447)

[References 12](#_Toc55919448)

# Figures

[Figure 1: Data Map 4](#_Toc55915869)

[Figure 2: Types of loan 7](#_Toc55915870)

[Figure 3: Loan size distribution before dealing with outliers 8](#_Toc55915871)

[Figure 4: Loan size distribution after dealing with outliers 8](#_Toc55915872)

[Figure 5: Income distribution after dealing with outliers 9](#_Toc55915873)

[Figure 6: Default count 9](#_Toc55915874)

[Figure 7: Default percentage of each type of loan 10](#_Toc55915875)

[Figure 8: The relation between loan size, income and target 10](#_Toc55915876)

# References

Akinwumi, J.A. and Ajayi, A. O. (1990). The Role of Loan Repayment Capacity in Small Scale Farmers. *Quarterly Journal of Credit Administration*, 15(2).

C. A. Wongnaa, D. Awunyo-Vitor . (2013). Factors Affecting Loan Repayment Performance Among Yam Farmers in the Sene District, Ghana. *Agris on-line Papers in Economics and Informatics*, 111-122.

Evans Brako Ntiamoah, Emmanuel Oteng, Beatrice Opoku, Anthony Siaw. (2014). Loan Default Rate and its Impact on Profitability in Financial Institutions. *Research Journal of Finance and Accounting*, 67-73.

Okpugie, G. (2009). High Microfinance Interest Rates Cause Loan Defaults in Nigeria. *The Guardian*.

Olomola, A. (2001). The nature and determinants of rural loan repayment performance in Nigeria: The case of FADU’s micro-credit programme. *NISER monograph series NO 3, (NISER)*.

Setargie, S. (2013). Credit Default Risk and its Determinants of Microfinance Industry in Ethiopia. *EJBE Vol. 3* .

Suraya Hanim Mokhtar, Gilbert Nartea, Christopher Gan. (n.d.). Determinants of microcredit loans repayment problem among microfinance borrowers in Malaysia. 33-45.

Vandel, K. (1993). A perspective on Mortgage Default Research. Journal of the American Real Estate and Urban Economics. Vol 23.