KAYSUL ISLAM ANIK

PREDICTING LOAN DEFAULT IN P2P LENDING

A comparative analysis between fsQCA and logit model

Master's Thesis in Information Systems
Master's Programme in Governance of Digitalization
Supervisor: József Mezei, Assistant Professor,
Faculty of Social Sciences, Business and Economics
Åbo Akademi University
Åbo 2019

ABSTRACT

Subject: Information Systems	
Writer: Kaysul Islam Anik	
Title: Predicting Loan Default in P2P Lending: A comparative analysis between fsQCA and logit model	
Supervisor: Dr. József Mezei	**Supervisor:**

Abstract:

Successful identification of the potential contributors that might lead to the outcome of loan status is a significant concern in peer-to-peer (P2P) lending system. As a part of risk management, P2P platforms attempts to keep the risk levels proportional to the expected returns. Determining the risk level for the lending process requires efficient prediction methods. As a measure of prediction, different statistical tools and intelligent computing tools might serve the purpose for the lenders. But, choosing the best tool that lead to efficient prediction of the outcome is always a challenge for the them. Moreover, with the onset of technological advancement, big data also adds to the problem by making the data asymmetric, complex, and unstructured. In order to address these issues, this research aims to focus on the dichotomous property of the crisp set with the qualitative comparative analysis. As a part of this, among many intelligent computing systems, fuzzy set qualitative comparative analysis (fsQCA) is tested with statistical logistic regression. This study covered the basic theoretical backgrounds of different methods and the terminologies of P2P lending platform as a starting point. Primarily this research compared the outcome from these models and discovered that fsQCA performs slightly better than the logistic regression for prediction model development. Additionally, fsQCA adds more explainable alternatives for the financial managers which can help them to understand the interplay of the conditions towards the outcome. Conditions that can lead to the outcome of loan default are discussed where fsQCA performed successfully with higher prediction score than the default ratio of the original dataset. Compared to the default prediction model, fsQCA performed better at developing non-default prediction model in this research. For this, separate analysis is conducted based on separate outcome since the analysis also showed that fsQCA results are non-inversible to predict the occurrence of opposite outcome. Based on the outcome of fsQCA and its property of linguistic terms, it provides more opportunity to select among the best suitable alternatives that can lead to the same outcome. It is observed to be a standard method which can handle outliers, deal with complex data and utilize mixed data. This research concludes that, fsQCA can be a potential tool of risk management for the financial managers where they need to balance between different alternatives based on real-life scenarios.

Keywords: p2p, lending club, fuzzy, fuzzy set, fsQCA, logit model, logistic regression, big data	
Date: 21.05.2019	**Number of pages: 82**

Table of Contents

List of Figures

List of Tables

1 CHAPTER 1: INTRODUCTION

Loan default occurs when a debtor fails to fulfill the obligations agreed as per the debt contract (Student loan default in the United States, 2018). More specifically, a default occurs when a borrower misses 3 installments within a 24-month period (Pearson, Jr. & Greeff, 2006). The same authors also termed this as a "risk threshold" for the lending institutions since loan default brings financial losses for those organizations. The chance of lending institutions towards facing loan default increases with the amount of disbursed loan. In general, this amount of loan is directly involved with the capital requirement. A lending institution might go for raising capital through credit if the associated risk is low compared to the existing loan. More often traditional financing institutions bear the crisis and risk of proper credit/loan risk management. This, credit risk always bears the probability of a decline in loan's value (sometimes being worthless) which is central to the health of any financial institution (Tsorhe, Aboagye, & Kyereboah-Coleman, 2019). This is often the major types of risk that financial institutions have to bear.

Many researchers even compared financial institutions (e.g. banks) to the blood arteries of human body which in this case, play a vital role for the economic growth (Bourke & Shanmugan, 1990). These financial institutions (e.g. banks) mostly face three different types of risk: financial risk (which includes credit risk), operational risk, and strategic risk (Saunders & Cornett, 1999). Bruett (2004) suggested a periodical review of the credit risk assessment tools to properly assess financial, operational and strategic risks. Mismanagement of any of these risks might lead the lending organization towards bankruptcy also (Bruett, Alternative Credit Technologies, LLC, Echange LLC, & Enterprise Solutions Global Consulting, 2004). Along with the threat of being bankrupt, lending institutions also bear many other costs. These costs might be from different delinquency situations including lost interest, opportunity cost of principal, legal fees and related costs (Baku & Smith, 2010). Therefore, credit risk (because of loan default) is inevitable for the lending institutions. Because of the credit risk, higher overhead costs, collateral and many other costs, these traditional financial institutions have to endure the incidence of frequent credit default. Hereby, Lending institutions are more concerned about the financial history and status of the applicant. Since the period of the loan might be longer, the ability of the applicant to repay the installments and the loan plays a crucial role in terms of investment decisions. For different types of applicants, the scenario might

be different. As the lending institutions rely a lot on the returns from proper investments which directly influences the profitability, they need to decide the appropriate mix of conditions which can satisfy both the investors and the debtors. Without determining the risk of loan default or loan delinquency a debtor carries, the lending institutions might not be able to set proper terms and conditions for that investment. Thus, it necessitates more strict measures and policy for credit management which impose higher rate on the borrowers. Contrarily, borrowers seek for credit with minimum rates, sometimes collateral-free loan as well. This inverse demand is predominant in the financial market which always necessitate an alternative option for solving the puzzle. Since the smooth monetary flow is important despite of different risks, lenders need a proper tool that satisfies the borrowers with lower thresholds and lenders with less risky investment. P2P (Peer-to-Peer) lending might help with lower threshold for the loan but the lender's perspective is still to be considered. Thus, a tool that can predict a debtor's standing in terms of possible loan default might ease the decision-making process. Different lending organizations follow different algorithms to predict the probability of loan default. This helps to compare with the standard benchmark and take the decision whether to grant the loan or not. This research focuses on predicting loan-default and non-default based on the data retrieved from a P2P platform. The study tries to distinguish the key differentiating factors of traditional statistical tools from more advanced intelligent systems.

1.1 Problem Statement

Investors must determine a standard level of investment in loan since it's related to the profitability. Low level of loan might indicate lazy capital and excessive allocation of capital in the form of loan, might indicate more risks to the lender. More credit brings more chances of default, as total credit by a firm is significantly related to default (Bhimani, Gulamhussen, & Lopes, 2013). Few other reasons that place consideration of loan default on top of the priority are the relationship of accurate prediction of loan default with the cost of the capital of the firms, global financial crisis in 2008 etc. (Bhimani, Gulamhussen, & Lopes, 2013). Successful prediction can give lending institutions prior clue about the financial distress that a borrower person/organization might face. Thus, it can help the lender to set investment terms and conditions according to the risk level (Predicting loan defaults, 2018). But the prediction is not easy for the decision makers due to the heterogeneity of the data being used. Most of the data available online can be

unstructured, complex, disorganized, and full of missing values etc. In real life, this is practical that all the information about the borrowers will not be available. These issues make the data processing far more complex for advanced algorithm also. Many researchers developed advanced predictive model which generally works better with smaller data sample. With the technological advancement, it is obvious that the data size and available information will be accessible through certain method. Thus, a proper method or tool that can handle the complexity of data and bring more meaningful result for the decision makers is important. Along with this, the real-life data are not always quantifiable but a mix of qualitative and quantitative. Different traditional methods are used to handle different types of data and which performs better separately. But a tool which can develop a method of handling both qualitative and quantitative information is very important. As a result, this research introduces an intelligent computing tool to compare the outcome with the traditional methods.

1.2 Research objective

Due to the fear of economic crisis and its association with the firm's profitability, loan default is being considered one of the prime issues. Because of its association with the profitability, institutions might sometimes ignore circumstances that might lead to loan default (contingencies) or ignore own standards just to compete in the market (competition). These contingencies and competition lie among the 5C's of bad credit as identified by Golden and Walker (1993) (Ntiamoah, Oteng, Opoku, & Siaw, 2014) . Recognizing the potential factors that might lead to loan default is a complex process which might always vary. As of traditional method, repayment ability was mostly determined by the financial ability and the asset held by an individual borrower (Paul, 2014). But different studies found different factors that might trigger the borrower's payment decision directly or sometimes indirectly. Borrowers financial history, characteristics, repayment willingness, demographics may also lead to loan default. It's important to properly identify the core factor where the investor want to put more weight before allocating the loan. This core factor can be based on more than one variable or a set of variables based on the types of the method an institution might choose.

To deal with loan default, different institutions try to adopt different parametric and non-parametric algorithms to identify patterns in large datasets and predict the outcome

(Bagherpour, 2018). There are many sophisticated tools in computational intelligence such as fuzzy systems, neural networks, support vector machines, rough sets, artificial immune systems, and evolutionary algorithms (Marqués, García, & Sánchez, 2012). Among many tools, one of the excellent tools that many researchers are suggesting those institutions to use is Fuzzy logic approach. From this fuzzy logic set, Qualitative Comparative Analysis using fuzzy sets (fsQCA) can be used to predict the probability of loan default from the dataset of new or existing customers. The aim is to develop a model which can be used for simulation and the outcome can indicate which loan tend to default and non-default based on the potential conditions. Based on the outcome, it could also be investigated if there is any specific set of variables which impact more on the status of the loan.

1.3 Research Questions

To fulfill the aim of this research, it is important to set the research questions. This research aims to find out the following questions-

RQ 1: Can fsQCA generate a better predictive model compared to Logistic Regression?

This research targets to use the collected data and run analysis for both logistic regression and fsQCA. The primary objective is to investigate if fsQCA can outperform the traditional method in terms of predictive model development.

RQ 2: Is it feasible to adopt fsQCA for the financial sector?

Traditional statistical tools are well accepted in most of the research fields. The aim of this research includes analyzing the data from financial sector. If the research question 1 performs successfully, it is to be deemed that fsQCA might have a prospect in the application of the method in business sectors also.

RQ 3: Best application of fsQCA: Predictive model for occurrence versus non-occurrence

In many cases, the output generated from traditional methods can be used to explain the reverse situation just by logically reversing or negating the output. It is to be investigated if the logical negation can be applied in fsQCA also.

RQ 4: Can fsQCA handle big data?

As many researches finds that many traditional models behave differently than the usual process in terms of larger data size, it is worth to investigate if the larger data size affects the output of the fsQCA analysis.

1.4 Scope of the research

This focus of this research is to develop a predictive model which can be used to predict the significant conditions that lead the loan to be default or non-default. This research focuses on the comparative analysis of traditional logistic regression with fsQCA method. As a workflow of this research, the aim is to conduct analysis for both of the methods and analyze the result to fulfill the research questions. For this purpose, a dataset of customers to whom loan is approved will be used and analyzed. Based on the output from the model, it might be pursued if there is any specific set of variables that mostly impact the consequences of loan status.

Relevant theoretical and mathematical background of this research is to be discussed for the ease of understanding the core knowledge of these methods. Background study about fuzzy logic, fsQCA, loan default, applications of fuzzy logic in different fields, logistic regression will be discussed also.

1.5 Structure of the research

For the purpose of the research, a theoretical review is preliminary conducted followed by the relevant analysis in different chapters. Chapter 1 describes the problem identification, formulation of the research objective, necessary research questions to be investigated using the research, the scope and the structure of the complete research. In chapter 2, the relevant theoretical and mathematical issues about peer-to-peer lending system, the fuzzy logic, fuzzy-set qualitative comparative analysis are discussed. Chapter 3 includes the literature review for the research. Since, implementation of fsQCA in business field is in the infancy level, literature review is conducted on a broader view including the basic fuzzy logic and fsQCA. Chapter 4 includes the descriptive settings of the variables and lead the way towards the analysis process. Chapter 5 includes traditional correlation, logit model and fsQCA analysis which are done for the purpose of further

analysis. It also includes the interpretation for the relevant methods separately. On, Chapter 6 a descriptive discussion based on different method is presented. The limitation of this study and fsQCA is also included in this chapter. Finally, Chapter 7 includes the conclusion and scope of future research in this relevant field. All the relevant references are cited following the APA format in the "Reference" section. The "Appendices" section also includes the contents which are considered necessary for the research.

To run the logistic regression and fsQCA a software called "R programming (Version 3.5.2)". For fsQCA a package titled "QCA" will be used. For reviewing literatures relevant to this research, a software called "NVivo 12" will be used. The whole report will be written using "Microsoft Word" and "Microsoft Visio" will be used for illustrating different diagrams and figures.

2 CHAPTER 2: THEORETICAL AND MATHEMATICAL DESCRIPTION

This chapter explains the theoretical and mathematical aspects relevant to this research. Since the scope of this research includes application of logistics regression analysis, fuzzy set qualitative comparative analysis (fsQCA) to predict the loan status in Peer-to-Peer lending, the basic terminologies of these different tools are covered in this chapter.

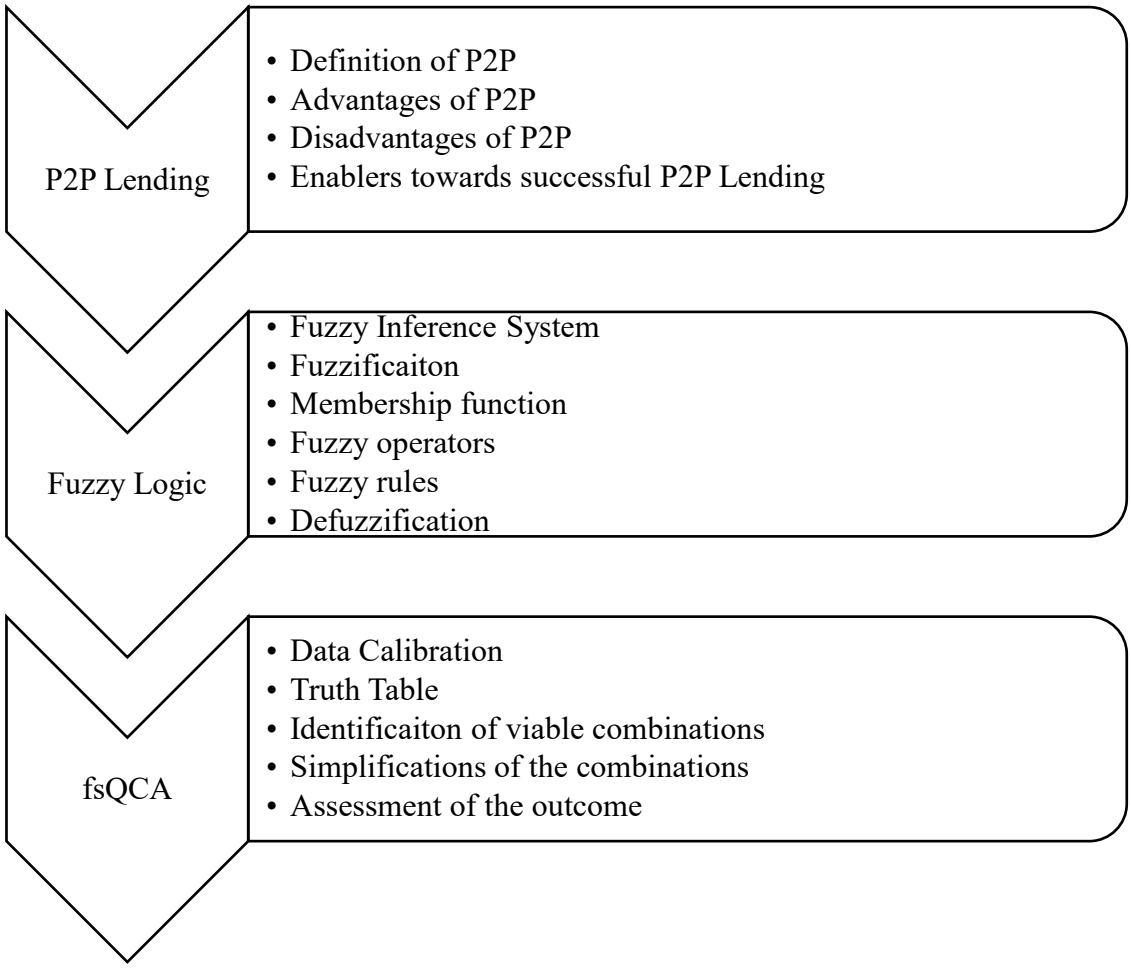


Figure 1 Outline of Chapter 2

2.1 P2P (Peer-to-Peer) lending

P2P lending is a process of lending money to individuals or businesses using different online services at a lower cost than the traditional financial institutions charge (Jiang, Wang, Wang, & Ding, 2018). More specifically, Peer-to-peer(P2P) is a person-to-person

lending that allows people to directly lend to and borrow from one another on an Internet-based platform without the participation of traditional financial intermediaries (Guo, Zhou, Luo, Liu, & Xiong, 2016). This participation includes the presence of a platform instead of traditional banking system allowing both the parties to interact with each other without the presence of any middleman. It offers lower access thresholds compared to traditional financial institutions. The borrower group can range from small and medium business owners, entrepreneurs, low-income earners, borrowers rejected by banks, individuals etc. (Jiang, Wang, Wang, & Ding, 2018). But, the crucial challenge for individual investors in P2P lending marketplaces is the effective distribution of their money across diverse loans by accurately assessing the credit risk of each loan (Guo, Zhou, Luo, Liu, & Xiong, 2016). Instead of these challenges, P2P is getting popular in these group of customers, P2P lending is developing rapidly in terms of volume of transactions and platforms.

According to a prediction by Transparency Market Research, the global P2P market is estimated to be worth US$897.85 billion by 2024 from US$26.16 billion in 2015 (TMR, 2019). The P2P market's revenue share is mostly concentrated in North America (about US$11.38 billion), Europe and Asia Pacific (in 2015). The same report also a tremendous growth which might increase by almost 48.2% within 2024. Some of the very well-known P2P platforms operating in the U.S.A and Europe are Lending Club Corporation (LC) (according to Zhao, et al., 2017, LC issued more than $ 24.6 billion as a loan in the year of 2016), Zopa, Prosper Marketplace, Upstart, Funding Circle, CircleBack Lending, Peerform, Pave, Daric, Borrowers First, SoFi, Ratesetter and Auxmoney (Bajpai, 2016). Among many other different P2P lending platforms some other popular P2P platforms are Prosper, Kiva and Renrendai (Zhao, et al., 2017).

2.1.1 Peer-to-peer Lending process

Unlike the traditional lending system, P2P lenders make direct investments on the lending website. They can get information about the online borrowers through the platform. This detailed information helps to eliminate information asymmetry among the trading parties. On the other hand, the borrowers can indicate credibility through different functions of the platform. This assists the lender to search loan request, make comparisons and take a lending decision (Wang, Chen, Zhu, & Song, 2015). They have also developed a

schematic work flow of the P2P lending process which gives a clear idea how this system work between different parties.

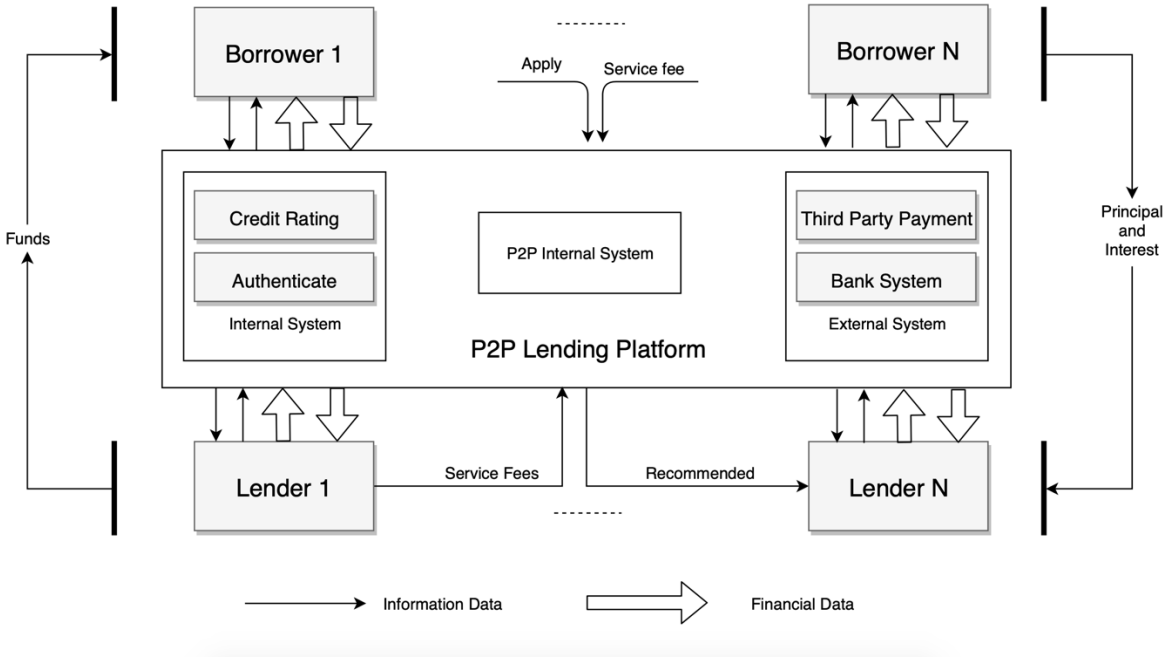


Figure 2 Peer-to-peer (P2P) lending process

This shows how number of lenders and borrowers interact with each other through a P2P lending platform.

2.1.2 Advantages of P2P lending

With the increase in sophisticated technology the cloud computing has become a very popular nowadays. With the outburst of cloud computing, web 2.0, big data, platform business has become a highly competitive tool compared to the traditional value chain system. Moore's law also summarizes the increase in processing power which brings newer technology along with substantial number of users. With the increase in these technologies, p2p has become a rapid-growing market that grabs attentions of many users (borrowers and lenders) and generates massive transaction data (Zhao, et al., 2017). This work as an interaction point for both of the parties for fulfilling their own agenda.

Wang et al., (2015) compared traditional banking system to p2p lending. They pointed out that the traditional lending system includes many steps prior to deciding the lending decision whereby p2p tries to remove the need of middleman. This on the other hand reduce the extra time, extra cost for the transaction. Peer-to-peer (P2P) lending eliminates

the need for a traditional financial intermediary that results the low financing costs (Xia, Liu, & Liu, 2017).

The online peer-to-peer (P2P) lending market is such a platform where borrowers and lenders can meet virtually and implement loan business without the help of any traditional financial institutions (Lin, Li, & Zheng, 2017). It helps to make the transaction more secured, transparent and geographically accessible.

Unlike the traditional banking systems, P2P does not allow intermediary parties. It helps both the lenders and borrowers to reduce the overhead cost. Since, P2P lending platforms can run with low overhead because of its exclusive online operation it enables borrowers to obtain loan at a relatively low interest rate (Xia, Liu, & Liu, 2017)

Since most of the P2P lending platform works specially based on the credit standing of the borrowers, it's important that the borrower construct a good credit score as to receive loan. In the traditional banks, it is imperative that the borrower need to fulfill many formalities where many borrowers might fail to ascertain a good score. Here, in this case, P2P lending fascinates number of borrowers who have lack of credit status and are unqualified for traditional bank loans (Xia, Liu, & Liu, 2017).

An Internet-based platform, P2P lending has been familiarized as a new e-commerce phenomenon in the financial field and it has great potential to provide more economical efficiencies (Lin, Li, & Zheng, 2017).

2.1.3 Disadvantages of P2P lending

Due to the very nature of the P2P lending system, the borrowers and lenders are not known to each other. The system works on the basis of platform and credit standing. This creates an information asymmetry between borrowers and lenders which might create problem in online P2P lending market (Lin, Li, & Zheng, 2017). This can also lead to trust issues among the trading parties.

Since the platform is accumulating number of lenders with number of borrowers, it's the responsibility of the lenders to check the eligibility and credibility of the borrower prior to making the lending decision. But this is very common that, lenders lack financial expertise and professional skills to realize potential high-risk borrowers (Xia, Liu, & Liu, 2017).

In most of the traditional banks, they often require collateral for the loan. But P2P offers collateral free loan to the borrowers. This makes the P2P loans typically unsecured and as a result, lender seek higher returns for the financial risk they are bearing (Xia, Liu, & Liu, 2017). This high return might sometimes overcome the advantage of low overhead cost that the platform offers for the borrowers. Thus, the ultimate interest rate can be higher due to excessive financial risk.

2.1.4 Enablers towards successful P2P lending

An article based on a large P2P lending platform data in China has revealed some influential demographic characteristics of borrowers that determine the default risk. The empirical result discloses that gender, age, marital status, educational level, working years, company size, monthly payment, loan amount, debt to income ratio and delinquency history play an important role in loan defaults. The research shows that there are many characteristics of borrowers that are related with low default risk (Lin, Li, & Zheng, 2017). These characteristics are shortly explained here.

Gender

Men and women react differently towards any borrowing decision. Generally, women have less risk seeking behaviors in compared to man. The research report of china shows that women have relatively lower default rate (Lin, Li, & Zheng, 2017).

Young adults

The default rate of a loan decreases with the increase of the borrower's working age. At the same time, the default rate increases with the increase of borrower's age. The older a borrower gets, the heavier the burden of his family is, so the higher his/her default risk will be (Lin, Li, & Zheng, 2017).

Low working time

The more years a borrower worked for, the better his financial status will be, and meanwhile people who work for a long time are more likely to realize the importance of credit (Lin, Li, & Zheng, 2017).

Stable marital status

Borrowers who have normal and healthy marriage have lower default risk, which is in accord with our normal cognitive. Divorced borrowers need to take care of their family alone, so they are in relatively poor financial situation, and they are likely to be overdue in the future (Lin, Li, & Zheng, 2017).

High education level

Highly educated borrowers, especially borrowers who have bachelor's degree or above, their default rate is significantly lower than those borrowers with low education degree, this suggests that an increase in the educational level (that is, more formal education received) will decrease the probability of loan default. The reason of this phenomenon, on the one hand, maybe people with higher education level are more aware of reputation and prestige, they are more likely to pay bills on time; On the other hand, it may be a person who understands the value of credit is more likely to get a higher degree (Lin, Li, & Zheng, 2017).

Employment in a large company

On the other hand, it may be a person who understands the value of credit is more likely to get a higher degree. The default rate of a borrower who works in large company is lower than that of a borrower who works in small company. The reason could be the one who works in large enterprise has more colleagues, the adverse effects of default is more serious for them, and also, the financial status of borrowers who work in large company is better (higher income, and they can borrow money from more colleagues), so their abilities of repaying timely is stronger (Lin, Li, & Zheng, 2017).

Low debt to income ratio (dti)

The possible reason for this funny phenomenon is that borrowers whose debt-to-income ratio is low may not care about the economic cost of default, and the cost of overdue payment is relatively small (Lin, Li, & Zheng, 2017).

No default history/ Delinquency history

If a borrower had default behavior at one time, his probability of loan default will increase dramatically. For lenders, this finding implies that lenders should not lend money to the

ones who have delinquency history; for borrowers, it reveals the importance of credit and the large cost of loan default (Lin, Li, & Zheng, 2017).

2.2 The Fuzzy Logic

Fuzzy logic is based on approximate reasoning which includes "degree of truth" whereby Boolean logic or Crisp logic includes "0" or "1". It was used by Lotfi Zadeh in 1965 which is an extension of Boolean logic and classical set theory. It engages human reasoning unlike the classical logic. In other words, it allows to include partial membership using values ranging from 0 to 1, keeping the core properties of set relation (Rihoux & Ragin, 2009). In practical life, vagueness of uncertainty always exists. Based on the properties of fuzzy sets and by analyzing and including expert opinions it helps to understand and model difficult real-life situations. It's been used in many fields which require modeling and control systems. Industrial control, human decision making, image processing is a few of them (Bennouna & Tkiouat, 2018).

2.2.1 Fuzzy Inference System (FIS)

It uses the properties of fuzzy set theory to convert the input space to an outer space. In this process fuzzy membership functions are used rather than using traditional Boolean logic. It's a process of attaching the changing environment to the computational model through expert knowledge (Chaudhari & Patil, 2014). The most common model used by the researchers is Mamdani Fuzzy (Özari & Ulusoy, 2017), (Bennouna & Tkiouat, 2018). Other FIS includes Sugeno and Tsukamoto fuzzy inference system (Chaudhari & Patil, 2014). This algorithm includes three major steps: fuzzification, inference and defuzzification. Chaudhari & Patil (2014) also mentioned about three core components of FIS which are 'Rule Base', 'Database' and 'Reasoning Mechanism'. 'Rules' base selects fuzzy rules which are in "IF…THEN" form. 'Database' defines the membership functions to be used in fuzzy rules. And 'Reasoning Mechanism' does the inference procedure based on the input information and fuzzy rules to conclude the output.

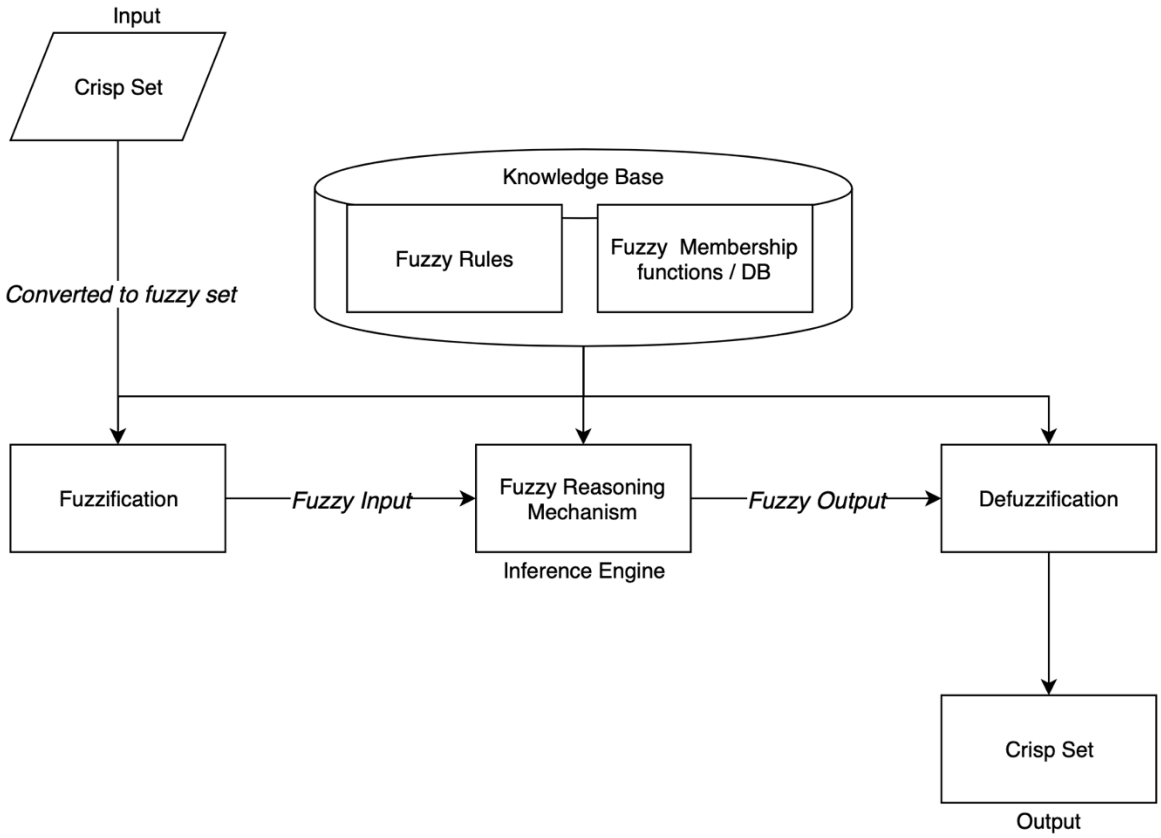


Figure 3 Fuzzy Inference System (FIS)

2.2.2 Fuzzification

Fuzzification converts crisp set to fuzzy set normally and as a further step, converts fuzzy set to a fuzzier set again. Thus, the linguistic variables of the fuzzy process are adjusted for processing. Different researchers might follow different methods which can be 'Support Fuzzification' or 'Grade Fuzzification'.

2.2.3 Membership functions

The fuzziness in fuzzy logic process is described by membership function which describes the degree of truth. Membership functions include three primary parts that include core, support and boundary. The core part consists of a full membership in the set, support part includes a region with non-zero membership in the set, and the boundary consists of non-zero but incomplete membership in the set. This membership allows to represent a fuzzy set graphically where universe of discourse is plotted in x-axis and the degree of membership in y-axis.

Different types of membership functions can be used by different researchers. That depends on the type and interest of the study being investigated. Triangular-function is one of the membership functions which has a value in between the upper and lower limit. Another type of membership function includes, Trapezoidal-function. Similar to Triangular-function, it has a lower and upper limit but both of these edge limit has separate support limit for each of them. There are R-functions and L-functions in Trapezoidal-function. Bennouna & Tkiouat (2018) used this Trapezoidal-function for their research because of the simplicity of this function (Bennouna & Tkiouat, 2018). Another membership function is Gaussian-function which looks bell shaped. It consists of a central value and Standard Deviation (SD). The smaller the SD is the narrower the bell-shaped figure is.

2.2.4 Fuzzy Operators

Ragin (2008) emphasized three common operations for fuzzy sets. The first one is: "Negation"; second one is: "logical AND" and the final one is: "logical OR" are important operations for fuzzy sets.

Negation

Normally in crisp set the membership changes its value from 0 to 1 and vice versa for negation. The same rules also apply for fuzzy sets. To calculate the negation of fuzzy set, the membership value should be deducted from 1. The negation value of fuzzy set is denoted by tilde sign (~). For a set X, if we want to calculate membership not in X, it the membership has to be deducted from 1.

$\sim X = [1] - X$

Logical AND

It uses 'set interaction' to combine two or more sets and takes the minimum membership of each score of each case from the combined sets. It uses the 'weakest link reasoning', which selects the lowest membership score to provide the degree of membership. This means that, if more sets are added to a combination of conditions, the membership score can stay same or decrease.

Logical OR

Unlike "Logical AND" operations, it uses union of sets and takes the maximum membership of each score of each case to combine two or more sets.

2.2.5 Fuzzy rules

Fuzzy logic can map the input space to an output space using IF…THEN rules. Mammadli (2016) used "IF…Then..." rules to formulate the model. Also, more complex rules along with higher number of linguistic variables were recommended for more realistic results (Mammadli, 2016). Menekay (2016) also followed IF...THEN rules to construct the knowledge base. But the author put more focus on the knowledge base of the experts (Menekay, 2016).

Malhotra & Malhotra (2002) mentioned that the fuzzy inference system works in five steps: fuzzification of the inputs, apply fuzzy indicator, apply implication method, aggregate all outputs, and defuzzification the output (Malhotra & Malhotra, 2002).

As of the fuzzy logic process, developing membership function is one of the most important tasks. It can be based on the knowledge of the experts or through some advanced algorithms. Tsabadze (2017) developed the membership function and the corresponding elements of weight vector based on intuition and knowledge base of the experts (Tsabadze, 2017). To develop the index, Özari & Ulusoy (2017) used factor analysis to eliminate some of them which was followed by clustering. The idea was to find out the volatile sectors (the most important sectors that remain in different clusters) and decrease the number of input variables to make the index more accurate. For these variables similar number of membership functions were developed. In the fuzzy logic process, they used and/or method. For the implication process, they took minimum and for the aggregation process they used maximum method. For the defuzzification, they used centroid method. Finally, for the output, Mamdani fuzzy inference system was used (Özari & Ulusoy, 2017).

Bennouna & Tkiouat (2018) categorized input variables were into three main categories and then used Mamdani Model for fuzzification. Finally, the developed model was used to simulate a set of customers to predict the customer behavior. (Bennouna & Tkiouat, 2018).

2.2.6 Defuzzification

Since the fuzzification process converts crisp sets to fuzzy sets for processing, defuzzification converts fuzzy set to a crisp set. Based on the necessity of the research, this process is carried out. Different methods are followed by different researchers. Most popular methods are 'Centroid Method', 'Max-membership method', 'Weighted Average Method', and 'Mean-Max Membership' etc. Centroid method is used because of its simplicity and short calculation time (Bennouna & Tkiouat, 2018), (Khan & Haque, 2013), (Chaudhari & Patil, 2014).

2.3 Fuzzy-set Qualitative Comparative Analysis (fsQCA)

Based on fuzzy set theory, it relates to different classes of objects having non-sharp boundaries where membership is a measure of degree (Zadeh, 1995). Ragin (2008) first introduced fsQCA. It uses Boolean algebra and algorithms to reduce a large number of complex causal conditions to a small set of configurations that lead to a certain outcome (Felício, Duarte, & Rodrigues, 2016). There is a significant difference between fsQCA and other conventional statistical techniques (Malhotra & Malhotra, 2002); (Tsabadze, 2017). Those traditional methods measure the effect on the outcome whether fsQCA tries to find the conditions behind a given solution (Schneider & Wagemann, 2010); (Chari, Tarkiainen, & Salojärvi, 2016). To overcome the problems faced by the conventional methods, it helps to deal with two different issues: asymmetry and non-linear relationship (Chari, Tarkiainen, & Salojärvi, 2016). Apart from dealing with these, it deals with equifinality (different combinations lead to the same outcome) and casual complexity (combination of casual antecedents with core factors) (Elliott, 2013). Thus, this combination of the plausible outcomes can provide more scope of analysis for the researchers. To analyze the data using fsQCA, researchers mainly follow the following steps of fsQCA process.

2.3.1 Data calibration

The first step of fsQCA is to calibrate all the variables into sets which represents the degree of membership where a predictor variable takes place in a category. As a fsQCA process, all the variables are converted to a set. To be more specific, this set is not a variable but a group of values that displays the degree of membership in a specific

condition or category (Woodside & Zhang, 2013). According to the properties set theory, the values in a set can be either 0 or 1 (crisp set) but the fuzzy set entail varying degree of membership that ranges from 0 to 1 (Skarmeas, Leonidou, & Saridakis, 2014). Within this range, fsQCA can usually take values of 0.05 (complete non-membership), 0.5 (complete ambiguity) and 0.95 (complete membership). Similarly, there can be four value fuzzy set consisting 1, 0.67 (more in than out), 0.33 (more out than in) and 0. Six value fuzzy set can contain 1, 0.9 (mostly but not fully in), 0.6 (more or less in), 0.4 (more or less out), 0.1 (mostly but not fully out), and 0. The continuous fuzzy set can contain values of 1, $0.5<Xi<1$ (more in than out), 0.5 (cross over), $0 <Xi<0.5$ (more out than in) and 0 (Ragin, 2008). Ragin (2008) also mentioned that this number of intervals solely depend at the discretion of the researcher. This is based on the substantial knowledge of the researchers.

Sometimes, researchers might also be interested in the negated sets to express the absence of a certain condition (Woodside & Zhang, 2013). That negated set can be calculating by deducting the membership from 1 (e.g. if the membership is denoted by A, the negated set will be denoted by ~A) (Skarmeas, Leonidou, & Saridakis, 2014).

2.3.2 The Truth Table

The second step is the formation of truth table. It identifies the combination of predictor variable with the outcome. It holds 2^k number of rows where k denotes the conditions that is used in the analysis process. Ragin (2008) mentioned that the rows of truth table represents all the possible combinations of causes with the outcome. In this table, the rows might have none to many cases.

2.3.3 Identification of viable combination

In this step, relevant combinations are sorted based on the criteria that the combinations have non-zero observation between the predictor and the outcome. The most challenging task is to minimize the number of combinations from the truth table. Minimum number of cases for the outcome and the consistency level can be considered during this process (Ragin, 2000). Also, “complex causal recipe” can be employed which uses “and” terminology and takes the minimum scores of the conditions (Skarmeas, Leonidou, & Saridakis, 2014).

2.3.4 Simplification of the combinations

The truth table holds a range of plausible solutions that includes core and peripheral combinations between causes and the outcome (Ragin, 2000). The algorithm used for this is called counterfactual analysis. The fsQCA might provide three different types of solutions: complex, parsimonious and intermediate (Rihoux & Ragin, 2009). In complex solution, no simplifying assumptions are considered. As a result, the solution is mostly complicated due to the large number of casual antecedent conditions. In order to overcome this complicated outcome, parsimonious solution can be used which uses the antecedent conditions which are not observed in the dataset. With a strong assumption it needs full justification prior to implementation. Similar to the parsimonious solution, the intermediate also considers just the easy remainders by distinguishing "easy" and "strong" assumptions (Skarmeas, Leonidou, & Saridakis, 2014).

2.3.5 Assessment of the outcome

The final task is to assess the solutions based on the consistency and the coverage. The consistency represents the extent to which a casual combination leads to outcome (Skarmeas, Leonidou, & Saridakis, 2014). Normally, it is measured using the predetermined threshold. Different researchers use different threshold based on the types of the analysis. A consistency threshold of at least 0.75 to 0.95 is recommended (Ragin, 2000). But, Skarmeas et al., (2014) mentioned the cutoff threshold to be equal or more than 0.80. Although there have been different benchmarks for the thresholds, but the selection of appropriated threshold completely depends on the type of research and the researcher. In the final stage, all the combinations above the selected consistency threshold is selected. These combinations with high consistency threshold (pathways) mostly lead to the given output (Elliott, 2013).

Coverage shows how many cases having high membership in the outcome condition are represented by a particular condition (Skarmeas, Leonidou, & Saridakis, 2014). A value ranging from 0.25 to 0.65 is considered to be a standard one (Rihoux & Ragin, 2009). Simply, if the consistency threshold is higher (the higher the final consistency will be), the coverage should be lower for the solution (Skarmeas, Leonidou, & Saridakis, 2014) (Elliott, 2013).

3 CHAPTER 3: LITERATURE REVIEW

This chapter includes a detailed literature review on fuzzy logic and the fuzzy set Qualitative Comparative Analysis (fsQCA). Since fsQCA is comparatively a new topic in business sector, its implementation in different fields along with financial sector is described.

3.1 Fuzzy logic

One of the key issues that lending institutions need to deal with is balancing the risk of loan allocation. An effective task to sort out the probable default in loan allocation might be to find out the risky customers. Mostly different organizations use scoring method for this purpose (Tsabadze, 2017). This is a technique used to estimate whether an applicant is solvent and can repay the loan (Menekay, 2016). Among the tools that institutions use, credit scoring is one of the mostly adopted. Different factors are generally considered (based on the experience) for the assessment in this type of analysis. It can be through statistical model which might lack the history of loan default handling or theoretical model which might lack the information on stock exchange (Tsabadze, 2017). Malhotra & Malhotra (2002) also discussed about different traditional tools such as statistical models, credit scoring models or experience-based rules. For example, credit scoring models can never eliminate human element from the model. Thus, subjective decision always impacts the decision-making process for the scores which is in between accept-scores and reject-scores. On one hand, proper analysis must aim to minimize subjective judgement or human factor. On the other hand, a good predictive model necessitates a system to be capable of behaving like human brains (Malhotra & Malhotra, 2002). As a desperate need to this, researchers always try to develop an effective forecasting model (Tsabadze, 2017). According to many researchers, a tentative solution to these types of analysis could be the use of artificial intelligence applications which are already being used in financial services industry (Menekay, 2016). Among many of these, fuzzy system has attracted the growing interest of many researchers and practitioners (Malhotra & Malhotra, 2002).

Abdulrahman et al., (2014) on their study used fuzzy logic to form a model which combined both statistical and the subjective judgement together for decision making. This

model minimized human efforts challenges (also described in 5Cs of bad credit by Golden and Walker, 1993) that lending institutions face with the loan allocation (Abdulrahman, Panford, & Hayfron-Acquah, 2014). Similarly, Malhotra & Malhotra (2002), mentioned that fuzzy logic can deal with inexact information, blend the system with human experience. They also termed this 'inexact information' to 'non-sharp boundaries'. Bennouna & Tkiouat (2018), concluded in their research that, the fuzzy logic-based model included both statistical analysis and the degree of truth from the managers. Authors also suggested that the same model can be used for new or existing customers because of the weight of the descriptive variable (Bennouna & Tkiouat, 2018). But Tsabadze (2017) designed a scoring system to distinguish customers bearing high or low risk profiles which focused on the knowledge base of the experts and ignored statistical data. That research skipped defining the benchmark for labelling good or bad borrower was not accomplished thus acceptance or rejection of a specific loan could not be determined (Tsabadze, 2017).

Bennouna & Tkiouat (2018) applied fuzzy logic approach for credit scoring of microfinance institutions in Morocco. The research focused on evaluating customer behavior to reduce loan default, ensure the viability and sustainability of the microfinance institutions (Bennouna & Tkiouat, 2018). Abdulrahman et al., (2014) also used fuzzy logic for credit scoring and identifying tentative loan default for the microfinance institutions in Ghana. The model developed was aimed at counteracting the challenges of loan default. Since, fuzzy logic includes degree of membership, it was suggested to adjust the variables with the change in the economy (Abdulrahman, Panford, & Hayfron-Acquah, 2014). Whereby, Bennouna & Tkiouat (2018) analyzed customer behavior, Mammadli (2016) checked applicant's credit standing through the developed fuzzy-logic model and evaluated the quality of the retail loan. As defined by the degree of membership, the output shows the credit behavior of the customers from the model (Mammadli, 2016).

For the credit on organizational levels, Özari & Ulusoy (2017) developed an index to find out the default probability of any company which is termed as Fuzzy-bankruptcy index. They mentioned that the subjective factors (e.g. managerial arrogance, fraud, managerial mistakes) might impact on the result. Thus, a combination of different variables was determined using the factor analysis, clustering and Merton Model (MPD) (Özari & Ulusoy, 2017). Menekay (2016) suggested that the elimination of subjective influence

might also ensure speed and accuracy and outcome the efficiency received based on the knowledge of the experts (Menekay, 2016).

In terms of big data, imprecision and uncertainty of information is very common in data analysis. Different researchers use different method to deal with that. To tackle uncertainty during the decision making, Menekay (2016) used fuzzy expert system (which is also based on knowledge and logical rules). The primary focus was to combine knowledge base (KB) with inference engine for credit analysis and authorization. The author put more focus on the KB which was developed through the knowledge of experienced specialists and then the system processed input data based on the KB to generate final output (Menekay, 2016). To deal with imprecise data and non-linear functions of arbitrary complexity, Malhotra & Malhotra (2002) used a combination of the fuzzy logic and the neural network. That neuro-fuzzy model minimized Type I error, showed higher classification accuracy and outperformed linear discriminant regression model in identifying loan defaulters. (Malhotra & Malhotra, 2002)

3.2 Fuzzy-set Qualitative Comparative Analysis (fsQCA)

Felício et al., (2016) used fsQCA to understand the association of loan quality with the governance mechanisms. Same method is also used to identify the factors that lead to good financial performance and high loan quality. In their research, negation on non-performing asset ratio is used to identify high loans quality. Based on the configurations of fsQCA developed in that study, authors showed different combinations of variables that can lead to high loan quality (Felício, Rodrigues, & Samagaio, 2016).

Felício et al., (2016) used fsQCA to analyze how individual and corporate global mindset relate to SMEs' internationalization behavior. We know, QCA is used to reveal the pattern that support the existence of casual relationship rather than proving casual relationship (Schneider & Wagemann, 2010). Thus, Felício et al., (2016) explored the presence or absence of global mindset attributes that can lead to internationalization. Whereby, conventional variable-based statistical methods might yield a single solution, fsQCA yields different combination of outcome for SMEs' internationalization. This lead to a broad interpretation scope of the SME internationalization and showed how different solutions can lead to the same result (Felício, Duarte, & Rodrigues, 2016). Poorkavoos et al., (2016) also used the same method in the same sector. But they examined the

conditions that might lead to higher level of innovation. They considered knowledge transfer networks and organization's capacity for their study. They mentioned that, it's not necessary for any organization to follow the only one or the best output. Rather they suggested this method which provides different combinations of output. Small organizations can choose the best option that fits best with the resources (Poorkavoos, Duan, Edwards, & Ramanathan, 2016).

Tóth et al., (2017) offered a Generic Membership Evaluation Template (GMET) to support the human factor for assigning fuzzy set values to conditions. One of the most important limitation in qualitative calibration is dealt with this framework. Since, involvement of experts is important for membership functions, this can improve the transparency of the process. To prove the usefulness of using fsQCA in the complex situations, their study illustrated ways in which customers can achieve attractiveness in the eyes of the supplier (Tóth, Henneberg, & Naudé, 2017).

For exploring the causes towards customer knowledge utilization, Chari et al., (2016) mentioned that only the core factors (key account management, CRM) might not be sufficient for the decision making. A combination of other peripheral antecedents (customer relationship orientation, top management involvement, and formalization) with the core factors is always important for decision making. The authors also highlighted the trade-off decisions between these conditions where fsQCA might be handy since the traditional methods put only certain output. More specifically, in real life, the relationship are mostly asymmetric and non-linear whereby traditional methods struggle to deal with these types of situation. Hereby, authors suggested to use fsQCA for multiple solution paths that can lead to optimal outcome rather than one predictor condition alone. This helps to deal with the trade-off decision making process (Chari, Tarkiainen, & Salojärvi, 2016).

Sjödin et al., (2016) used this method to suggest several key paths (rather than only one outcome) that firms may follow to achieve advanced service offerings. They also mentioned about the multiple key paths and handling asymmetry as few of the advantages of this approach (Sjödin;Parida;& Kohtamäki, 2016).

To find the factors which might lead to economic growth, Allen & Allen (2015) used fsQCA approach. Their study concluded that wage flexibility and co-operative labor

relations might lead to growth. In their research they mentioned that fsQCA can identify the potential patterns that might lead to the economic growth and can deal with 'casual complexity'. These are similar to 'multiple outcome' and 'asymmetry' as mentioned by many other researchers (Allen & Allen, 2015).

3.3 Applications of Fuzzy Logic And fsQCA

Skarmeas et al., (2014) mentioned that the application of fsQCA in business and management sector is fewer, but it is being used in many sectors of business research. The diversified use of fuzzy system and fsQCA are mostly visible now-a-days. Few of the important sub-fields are:

Mammadli (2016) used fuzzy logic model to indicate the credit standing from the perspective of retail loan. Author mentioned about the use of traditional statistical tools for the purpose of identifying any risk associated with the individual loan system. He mentioned about the discriminate analysis and logistic regression that banks used to prefer. These traditional methods usually process historical data and the decision is provided as a crisp set. This means that the outcome from these analyses provides either yes or no outcome for the decision makers. The key issue that concerned the author of using this approach was the incompleteness, imprecision and uncertainty of information in real life. These traditional models always face problems in terms of qualitative set of data. Another limitation is business research is the subjective judgement which might impact the decision-making process directly. This is always a key issue that need to be double checked in case of credit decision making. Professionals generally uses knowledge base in order to measure the qualitative data. This leaves more question on the transparency and automation of the process followed in traditional statistical models. But the use of fsQCA might be able to handle this subjective judgement using the linguistic values such as "high", "adequate", "low" etc. Author showed how a combination of categorical and numerical variables can be used for analyzing the credit standing of the borrowers. All the categorical variables are converted using linguistic term "Short/Medium/Long" and the numerical variables are converted using "Low/Medium/High". In order to analyze the variables, he used FuzzyTech Business software using the fuzzy rules "IF and THEN". Later, developed a model that can be used to evaluate the retail loan using imprecise knowledge or human subjective judgement.

The potential variables that the author found out were income level, credit history, character, collateral and employment history of the borrower. He also suggested to use more complex linguistic terms in real life to get more realistic results (Mammadli, 2016).

Korol (2019) compared three different types of forecasting model (statistical, soft computing methods, and theoretical models). The author clearly mentioned about the benefit of using soft computing methods for imprecisely defined problems, incomplete data, imprecision and uncertainty. In real life, it is practical that the collected data about the loan applicant might always lack some fields which needs advanced soft computing skills to handle. Opposed to the traditional statistical methods, fuzzy logic can interpret vague and ambiguous concepts. For example, the traditional model will determine a threshold which can define the risk of any specific company at a risk of bankruptcy, but fuzzy logic can introduce the term such as "high risk of bankruptcy" or "low risk of bankruptcy". In order to show the applicability of the fuzzy model, the author used demographic and financial variables of a set of borrowers to understand the financial situation. Based on many sub variables listed under these three variables, it was pointed out that education and status of the employment influence the output of the model (Korol, 2019).

Boratyńska & Grzegorzewska (2018) used fsQCA for bankruptcy prediction in agricultural entities and compared with classical quantitative methods. They pointed out that, for the conventional methods, combination of both qualitative and quantitative nature of the data might bring some limitations. But, fsQCA overcome this qualitative-quantitative limitations and leads to more detailed understanding of the conditions under which a certain outcome can occur. Compared to traditional methods, it offers several different casual expressions with own level of consistency and coverage. In the analysis conducted, the authors pointed out that illiquidity and lack of profitability of the agricultural entities lead to the outcome of bankruptcy. On the other hand, the traditional discriminant analysis focused on financial indicators only. It is obvious that, fsQCA offers more detailed understanding of the outcome compared to regression-based analysis (Boratyńska & Grzegorzewska, 2018).

In most of the cases, the linguistic terms are used in the application of fuzzy logic to handle the subjective judgement and uncertainty of the inputs. As Mammadli (2016) used linguistic terms to convert variables into membership function, Romaniuk & Hall (1992)

used learning mechanism for the model called FUZZNET system. The authors also mentioned that the system can generate stand-alone expert system by rule formation, learning or a combination of these two where the author preferred the learning method. One very interesting issue that the author pointed is the learning knowledge bases from scratch that gives a scope of working with less amount of data and simulate a similar situation for the knowledge base. This FUZZNET model includes three cells (input, output and hidden cells) where every cell has a bias associated with it on a real number scale. The cells are connected through links which have weight associated with them. Similar to other fuzzy set-based model, each cell can take value within the range of 0 to 1. Using the FUZZNET model, the authors pointed that mainly financial and personality basis can determine the creditworthiness of a borrower.

Feldman & Treleaven (1994) discussed about the use of intelligent systems in financial sector. They mentioned fuzzy logic, neural networks, genetic algorithms, expert systems and rule induction to be the mostly accepted tools to be considered as intelligent systems. One of the major benefits of using fuzzy logic is that it can handle imprecise data by using the knowledge bases in a production rule format. This introduce the transparency of the rules used by the system and include the option to judgmentally revise them. The other benefit is that domain experts can engage in setting up the rule base. Fuzzy logic being one of the intelligent systems can be used in credit evaluation, customer profiling, risk assessment, fraud detection, portfolio optimization, asset forecasting, economic modelling, sales forecasting, retail outlet location etc. (Feldman & Treleaven, 1994).

As mentioned by Feldman & Treleaven in 1994 about the use of intelligent systems in business sectors. With the passage of time, the use of fuzzy logic is increasing more day-by-day. The popularity of this method is mentioned by many researchers. Its use in different fields are also being explored extensively. Few of the sectors where fuzzy logic can be a competitive tool as compared to other traditional methods and intelligent systems-

Sectors	**Literatures**
Loan or Credit Quality	(Felício;Rodrigues;& Samagaio, 2016); (Chen & Chiou, 1999), (Romaniuk & Hall, 1992);

Loan default and credit risk	(Tsabadze, 2017), (Menekay, 2016), (Malhotra & Malhotra, 2002), (Abdulrahman;Panford;& Hayfron-Acquah, 2014);
Financial Management	(Córdova;Molina;& López, 2017);
Bankruptcy	(Özari & Ulusoy, 2017);
Entrepreneurship and Innovation	(Kraus;Soriano;& Schüssler, 2018), (Apetrei;Paniagua;& Sapena, 2016), (Coduras;Clemente;& Ruiz, 2016), (Lisboa;Skarmeas;& Saridakis, 2016), (Beynon;Jones;& Pickernell, 2016), (Huarng & Roig-Tierno, 2016);
Organizational Research	(Fiss, 2011);
Corporate sector	(Wang;Yu;& Chiang, 2016);
Performance Evaluation	(Mangaraj, 2016);
Customer Relationship Management	(Tóth, Henneberg, & Naudé, 2017), (Chari, Tarkiainen, & Salojärvi, 2016);
Service Offerings	(Sjödin;Parida;& Kohtamäki, 2016);
Customer Satisfaction	(Hsiao;Chen;Chang;& Chiu, 2016);
Small and Medium Enterprise Sector	(Felício, Duarte, & Rodrigues, 2016), (Poorkavoos, Duan, Edwards, & Ramanathan, 2016);
Microfinance	(Bennouna & Tkiouat, 2018);
Economic growth	(Allen & Allen, 2015);
International business	(Schneider & Wagemann, 2010);
Tax Legislation Reforms	(Musayev;Madatova;& Rustamov, 2016);
Energy Sector – Oil consumption	(Ramazanov;Jabiyeva;& Amirguliyev, 2016);
Lexicological dictionary	(Abdullayev;Umarova;Jamalov;& Alekperov, 2016);
Linguistics	(Mendel & Korjani, 2012)

Table 1 Application of fuzzy logic in different fields

4 CHAPTER 4: BUILDING PREDICTIVE MODEL (STAGE:1)

This chapter explains the preparation process of the dataset for the analysis. Based on the schematic flow of "Predictive Model Development" by Shmueli & Koppius (2011), the workflow has been divided into stage 1 and stage 2. In this chapter, stage 1 explains the overall objective of the research, data collection methodology, data transformation and exploratory data analysis for deeper understanding of the dataset. The methodology on how the necessary variables are selected based on the descriptive variables are also discussed in this section.

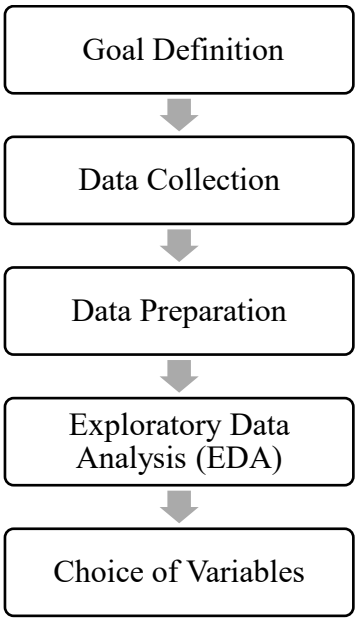


Figure 4 Outline of Chapter 4

Predictive Model

A model that produces highly informed guesses about any future outcome can be termed as predictive model. This highly informed guesses are not based on guesses only but some sophisticated algorithm. Different types of predictive model might have different baseline depending on the need of the outcome to be predicted. In this research, the primary target is to predict if a loan will result in default or non-default based on the user information. This does not just focus on the behavior but on the overall situation of the user. Many different scenarios need to be taken care of for a successful predictive model. This

necessitates a uniform and scientific process that can result in an accurate model development.

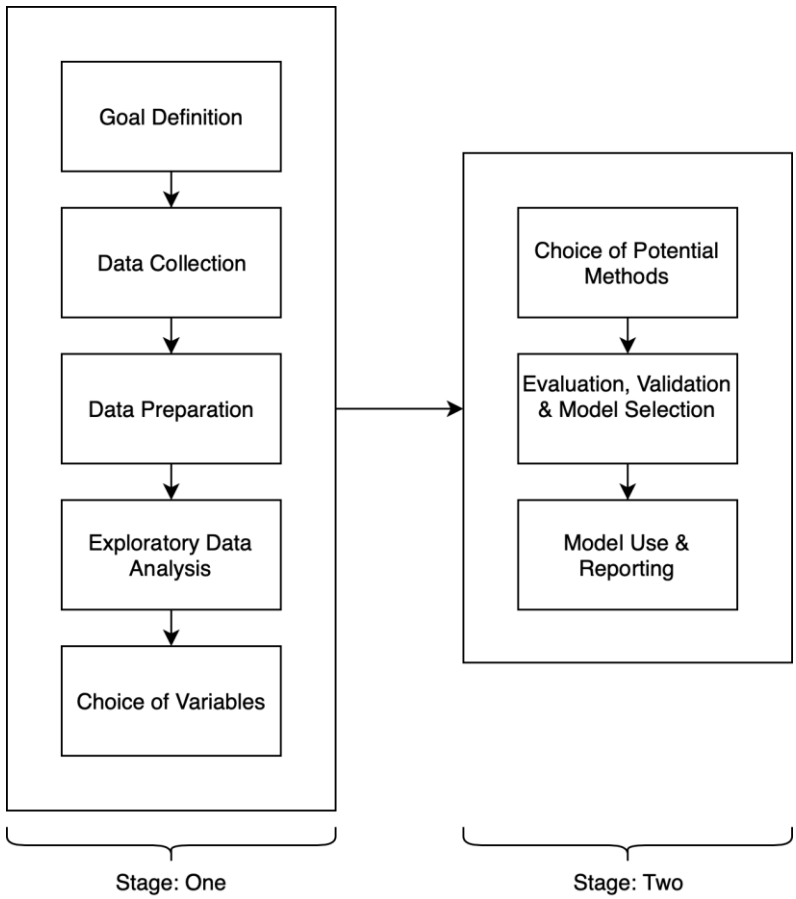


Figure 5 Schematic of the Steps in Building an Empirical Model (Predictive or Explanatory)

The above steps are based on the steps as mentioned by Shmueli & Koppius (2011). They mentioned a schematic of the model building steps for both explanatory and predictive modeling which includes eight important steps (Shmueli & Koppius, 2011). In this research, mostly these steps will be followed to build the predictive model. Compared to the original steps, the whole process is divided into two parts separating the analysis and interpretation parts from the descriptive statistics. Stage one from the above figure will be described in this chapter and stage two will be discussed in the following chapter.

4.1 Goal Definition

One of alarming issue of the digitalization is mass amount of data being generated each day. As we know that the raw data does not produce any explainable result, it is mandatory to process that vast amount of data in order to take a decision. Similarly,

information about almost every aspect of their life style is available online but due to lack of processing, those data never serves in a meaningful way. This creates a major problem for the businesspersons who needs to assess the user behavior for business needs. In this research, we tend to use big data and build a model that can indicate if a borrower sharing a certain type of characteristics may result in default or non-default. The strength or contribution of those certain characteristics which contribute to loan default might vary from small to high which normal statistical analysis fail to explain. Thus, this research aims to compare the predictability of the traditional statistical tool and explain together with the output based on advanced machine learning tool. Different tools like correlation, logistic regression, fsQCA analysis will be used in the analysis section.

4.2 Data collection

The data is retrieved from the official website of Lending Club. Lending club is a San Francisco, California based P2P lending company operating since 2007. It is a technology company which focuses on reinventing seamless credit at a low cost with high opportunity. It gathers both investors and borrowers in an online platform and helps investors to choose potential consumer credit asset class. To make the investment option safer it provides diversifying opportunity through different types of notes. According to the official website of Lending Club, 99% of the investors who held more than 100 notes (grade A to E), continued till the maturity and received positive returns. These different types of note terms correspond to the length of borrower loans, 36 or 60 months (3 or 5 years). The time period of these notes that corresponds to loan determines risk and return for the investors. The more maturity time it holds, the riskier and more return it provides (Lending Club, 2019). Simply lengthier the time period for a loan is, the chances of that loan being default becomes higher which bears more risk to the investment portfolio.

Data dimension

The dataset includes detailed records of the borrowers who applied for the loan and were accepted. The time period for the loan covers 2007 till the second quarter of 2018 with the geographical area of the borrowers covers different states of the U.S.A. The raw dataset includes a large dimension which captures various sources of information. Although all these variables can be used to discover new relationships, some of those variables will be selected based on the domain knowledge.

Dimension reduction

From the large raw dimensions of the dataset about 10 initial variables are selected for preliminary analysis. The selection was based on the predictive goal and the subjective knowledge.

	Variable Name	**Variable Status**	**Variable Type**	**Variable Details**
1	term	Independent	Categorical	The number of payments on the loan. The values of term are in months. It can be either 36 or 60 months.
2	funded_amnt	Independent	Numeric	The total amount of loan committed at that point in time.
3	int_rate	Independent	Numeric	Interest Rate on the loan
4	installment	Independent	Numeric	The monthly payment owed by the borrower if the loan originates.
5	dti	Independent	Numeric	A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income.
6	revol_util	Independent	Numeric	This is a ratio. It shows the ratio of the usage of credit line compared to total available credit.
7	emp_length	Independent	Categorical	Employment length in years. It can hold values between 0 and 10. Here, for one year it is 0; for ten or more years it is 10.
8	purpose	Independent	Categorical	This variable includes the

				intent of the borrower for the loan. This is provided by the borrowers during the application.
9	home_ownership	Independent	Categorical	This shows the status of the home ownership a borrower entail.
10	loan_status	Dependent	Categorical	Current status of the loan

Table 2 Details of the variables

4.3 Data Preparation

The raw dataset includes 2004091 records of the borrowers and based on the domain knowledge approximately 10 (1 dependent and 9 independent variables) were selected for further analysis. Initially, the variables are 4 categorical types: “term”, “emp_length”, “purpose” and “home_ownership”. The numeric variables are “funded_amnt”, “int_rate”, “installment”, “dti” and “revol_util”. The dependent variable also categorical type which will be converted to binary for the purpose of further analysis.

Missing values

The initial summary of the dataset consisted of a number of missing and NA values in the dataset. Most of those values were from the categorical variables. The proportion of the missing values to the amount of original dataset was 6.30%. Here, instead of mutating missing values with dummy variables, those records were deleted which still replaced the sample size from n = 2004091 to n = 1877775. So, new dimension of the dataset is 1877775 x 10, after the removal of the missing values.

4.4 Exploratory Data Analysis (EDA)

To run the correlation analysis, categorical variables are converted to numerical values based on different levels. Numeric variables are kept same and used in the analysis. The levels for “term”, “purpose” and “home_ownership” are kept same. For “emp_length”, instead of keeping 11 levels, it is customized to 4 broad levels which contain all the records. The Dependent variable “loan_status” had initial levels consisting "Fully Paid",

"Charged Off", "Current", "Late (31-120 days)", "In Grace Period", "Late (16-30 days)", "Default", "Does not meet the credit policy. Status: Fully Paid", "Does not meet the credit policy. Status: Charged Off". Based on the Lending Club Data Dictionary, it shows that the levels with "Default" and "Charged Off" are treated as default loan. Thus, for further analysis, “loan_status” is mutated to binary variable while "Default", "Charged Off" are coded with 1 and other levels are coded with 0.

Variable Name	**Data Mutation**
term	There are two types of loan repayment terms available where “60 months” = 1; “36 months” = 2.
emp_length	This variable has 11 labels. Here, "< 1 year" and "1 year" = 1; "2 years", "3 years", "4 years", "5 years" = 2; "6 years"; "7 years", "8 years", "9 years" = 3; "10+ years" = 4;
purpose	"debt_consolidation" = 1; "credit_card" = 2; "small_business" = 3; "major_purchase" = 4; "house" = 5; "home_improvement" = 6; "renewable_energy" = 7; "car" = 8; "medical" = 9; "vacation" = 10; "wedding" = 11; "educational" = 12; "moving" = 13; "other" = 14
home_ownership	The data consisted RENT, OWN, MORTGAGE, ANY, OTHER and NONE. Here, “RENT” = 1; “MORTGAGE” = 2; “OWN” = 3 and “ANY” = 4; “OTHER” = 5; “NONE” = 6;
loan_status	There are about 9 inputs in loan_status of which “Default” and “Charged Off” = 1 (defaulted), All the other inputs are coded with 0.

Table 3 Data Mutation

But for the logistic regression, this mutation is not conducted since it can easily deal with multiple levels of the categorical variables. Moreover, it’s easier to analyze the result with the levels on the output from a logistic regression. But only exception is that, the variable “emp_length” has too many levels which are converted to 4 major categories similar to the way as in Data Mutation table.

4.4.1 Descriptive statistics of variables

The way to get insight of the information from a dataset is to summarize the data using the descriptive statistics. Different types of central tendency, variation and shape might be used to analyze the information. In most of the cases, for the numeric data mean and median might be a good choice of explaining central tendency where frequency of observations, mode can be used for nominal data types. As we know, predictive models are based on the association rather than causation between the dependent and independent variables, the descriptive statistics are used for better insight of the information. A quick view for the variables from the summary statistics of the dataset:

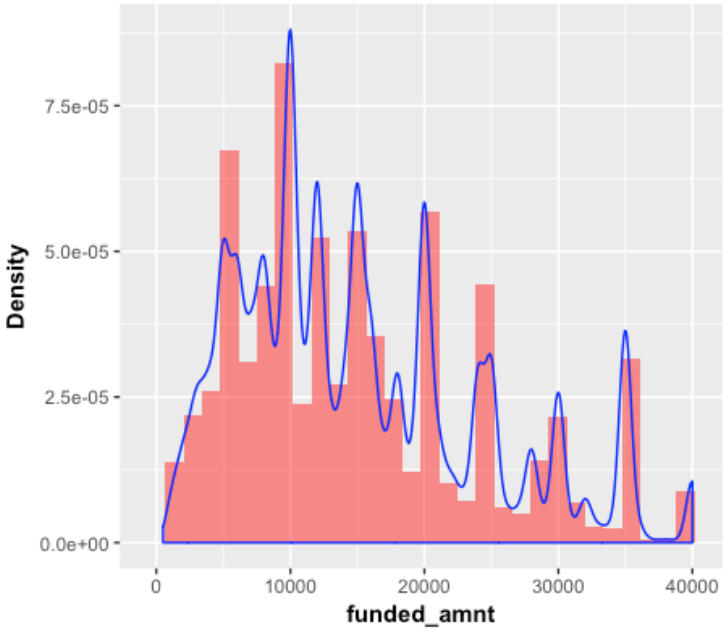
	term	funded_amnt	int_rate	installment	dti	revol_util	emp_length	purpose	home_ownership	loan_status
nbr.val	1877775	1877775	1877775	1877775	1877775	1877775	1877775	1877775	1877775	1877775
nbr.null	0	0	0	0	1100	10006	0	0	0	1719220
nbr.na	0	0	0	0	0	0	0	0	0	0
min	1	500	5.31	4.93	-1	0	1	1	1	0
max	2	40000	30.99	1719.83	999	892.3	4	14	6	1
range	1	39500	25.68	1714.9	1000	892.3	3	13	5	1
sum	3208783	28375926525	24613447.76	841711121.07	34703768.65	96487112.13	5127436	5268271	3203863	158555
median	2	13000	12.62	382.55	17.73	51.6	3	1	2	0
mean	1.71	15111.46	13.11	448.25	18.48	51.38	2.73	2.81	1.71	0.08
mode	2.00	10000.00	11.99	301.15	19.20	0.00	4.00	1.00	2.00	0.00
SE.mean	0	6.63	0	0.19	0.01	0.02	0	0	0	0
CI.mean.0.95	0	12.99	0.01	0.38	0.02	0.04	0	0.01	0	0
var	0.21	82472829.21	22.82	70420.35	120	602.76	1.22	12.39	0.42	0.08
std.dev	0.45	9081.46	4.78	265.37	10.95	24.55	1.1	3.52	0.65	0.28
coef.var	0.27	0.6	0.36	0.59	0.59	0.48	0.4	1.25	0.38	3.29
skewness	-0.92	0.75	0.76	0.99	20.7	-0.02	-0.15	2.3	0.4	2.99
kurtosis	-1.15	-0.16	0.65	0.68	1536.02	-0.07	-1.38	4.2	-0.6	6.94

Table 4 Descriptive statistics of the variables

4.5 Choice of Variables

Descriptive statistics show that all the variables which were selected initially can be used for the analysis. To get a more detailed view of the variables, those are graphically presented with various statistical measures.

funded_amnt – The total amount of loan committed at that point in time. In the original dataset, there was another variable named "loan_amnt", which included the amount a borrower applied for. That amount was not the final amount that the borrower was provided. Thus "funded_amnt" was used to find the exact amount the borrower was provided. Here, for the n number of records this the funded amount ranges from minimum USD 500 to 40000.

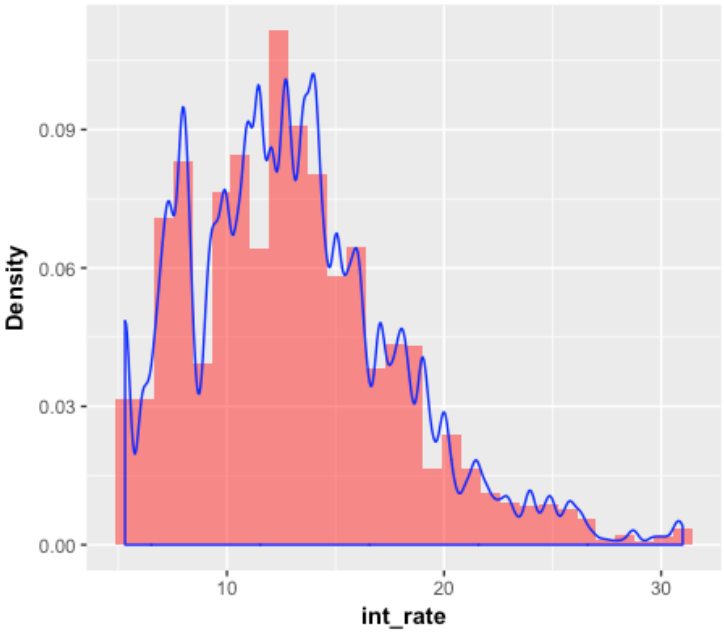


The mean value is 15111.46 which shows the average of the funded amount to the borrowers. As we know that for the skewed data mean value might not represent the center of the data. The median of 13000 shows that the data is a bit skewed but not much since the median is close to the mean.

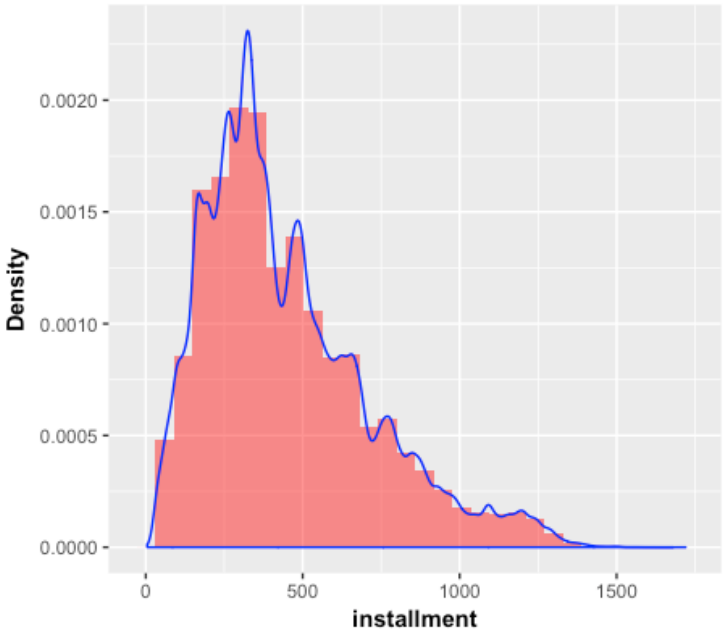
A mostly used measures of variation is Standard Deviation which measures the closeness of the observations to the mean. For a normally distributed data, $\mu \pm 1\sigma = 68\%$; $\mu \pm 2\sigma = 95\%$; $\mu \pm 3\sigma = 99.7\%$; where μ represents the mean and σ represents the Standard Deviation. For this variable the high std.dev of 9081.46 and mean of 15111.56 means that the data is quite spread from the mean value.

The skewness of the "funded_amnt" is 0.75 which shows that the distribution is normal since the skewness is almost near to 0 but the positive value shows that the distribution slightly right tailed. The excess kurtosis measures the outliers in the data or in other words the tail shape of any variable. While the normal distribution has a value of 0, the kurtosis value of -0.16 indicates that the distribution is thin-tailed (platykurtic).

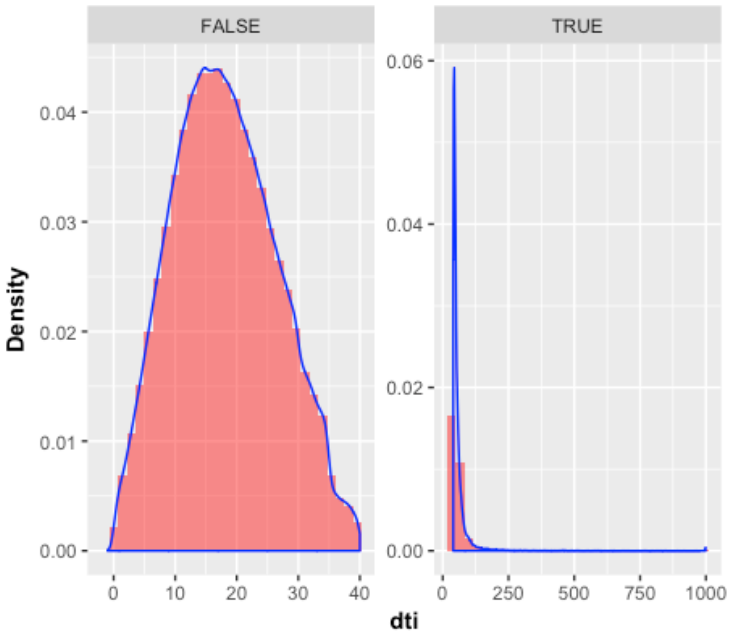
int_rate – This variable includes the interest rate applied to the borrowers for different category and amount of loans. The interest percentage ranges from 5.31 to 30.99 depending on the risk associated with the loan. Here, mean: 13.11; median: 12.62 and std.dev: 4.78 shows that the data is almost normally distributed and most of the value are close to the mean value. Skewness: 0.76 indicates that the distribution is normal and kurtosis: 0.65 indicates that the distribution is fat-tailed (leptokurtic).



installment - The monthly payment owed by the borrower for the amount of loan received. The amount of monthly installment ranges from 4.93 to 1719.83 depending on the amount and interest rate of the loan. Here, mean: 448.25; median: 382.55 and std.dev: 265.37 shows that the data is mostly normally distributed and most of the value are close to the mean value. Although the range of the installment amount is very high, the mean value shows that the average number of the amount of installment is low. Here, Skewness: 0.99 indicates that the distribution is mostly normal and kurtosis: 0.68 indicates that the distribution is fat-tailed (leptokurtic).



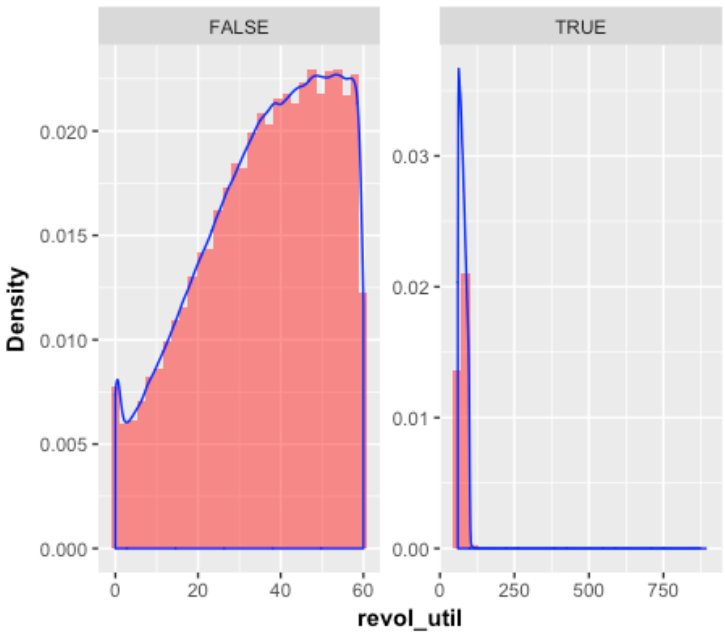
dti- This is a ratio determining the ability of the borrower to pay the total monthly debt from the income earned on that month. The ratio is calculated in the following way (Total monthly debt does not include the mortgage and requested loan from Lending Club):



$$dti = \frac{Total\ monthly\ debt\ obligations}{Total\ self\ reported\ monthly\ income}$$

While the value ranges from a minimum of -1 to 999, the lower the value is the more ability the borrower has to repay the loan. The mean and median value is almost identical and the std.dev: 10.95 shows that most of the data is close to the mean value. But, the kurtosis: 1536.02 shows that the distribution has outliers and as a result of this the mean value is also impacted.

revol_util – Revolving credit means a certain amount of available credit offered by a certain bank or lending institution to a borrower. "revol_util" is a ratio which shows the amount of used credit in comparison with the total available amount of revolving credit. The lower this ration is, the lower the amount of credit is used by that specific borrower. In a



word, the lower this number is the more flexible the terms and conditions is for the borrower leading to a better credit score.

$$revol\_util = \frac{Current\ usage\ of\ revolving\ credit}{Total\ available\ revolving\ credit}$$

The minimum value shows 0 which results because of the NULL values in the dataset. But the minimum values cannot be 0 in terms of a ratio variable. Thus, the range starts from a minimum value 0.01 to a maximum of 892.3. The median and the std.dev shows that most of the values are close to the average value. Skewness: -0.02 indicates that the distribution is normal and left tailed; kurtosis: -0.07 indicates that the distribution is thin-tailed (platykurtic).

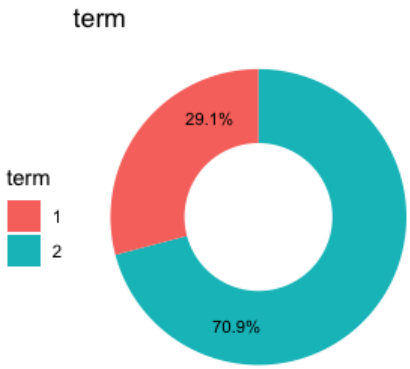
Frequency and proportion table for the categorical variables

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
term	546767	1331008												
%	29.1178	70.8822												
emp_length	293304	585927	331898	666646										
%	15.6198	31.2033	17.6751	35.5019										
Home_ownership	750835	928769	197465	482	177	47								
%	39.9854	49.4611	10.5159	0.0257	0.0094	0.0025								
purpose	1069925	418669	22040	42976	11645	124453	1221	20505	22310	12743	2321	412	13011	115544
%	56.9783	22.2960	1.1737	2.2887	0.6201	6.6277	0.065	1.0920	1.1881	0.6786	0.1236	0.0219	0.6929	6.1532

Table 5 Frequency and proportion table of the variables

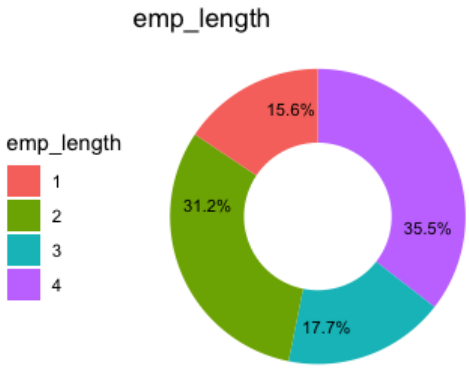
Here, the numbering in the columns shows the levels that each categorical variable has. The row showing % includes the percentage that each level for a specific variable holds based on the total n size of that variable. The cells which are highlighted shows the "Mode" of each variable.

term



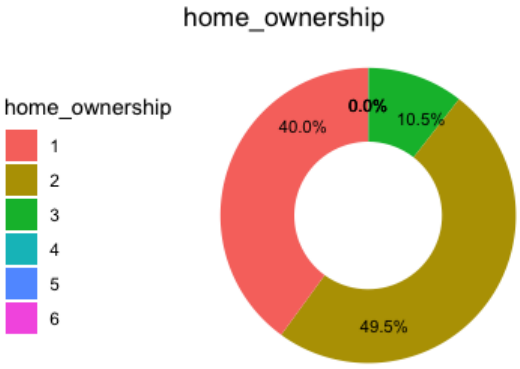
It shows the number of total payments to be done for the loan. According to the data available from Lending club, this value can be 60 months or 36 months. This data is mutated to numerical value 1 and 2. Here 1 consists of about 29.12% and 2 consists of 70.88% of the data. Level 2 also holds the mode for this variable.

emp_length



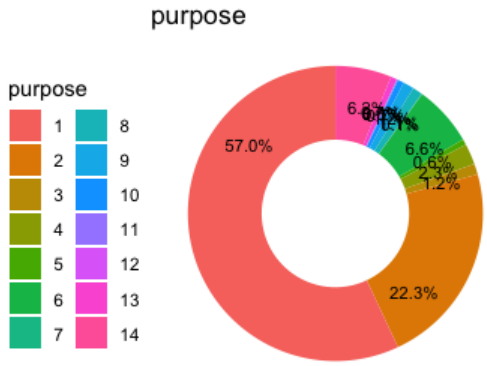
This shows the employment length of the borrowers. Data from Lending Club shows value between 0 to 10+ years. For the purpose of analysis, from 11 different levels, this variable is coded to 4 levels where 1: duration of 1 or less than 1 year; 2: duration of 2 to 5 years; 3: duration of 6 to 9 years; 4: 10 to more than 10 years of employment length. While the mode of this variable is in level 4 (35.50%) but level 2 also holds a significant amount of records (31.20%).

home_ownership



This shows the type of access a borrower has to home. The original dataset includes 6 different types of status which are rent, mortgage, own, any, other and none. All these types are converted to numerical values each having values from 1 to 6 simultaneously. Most significantly about 89.45% of the data is in the level 1 and level 2. The mode of this variable is in level 2 consisting of 49.46%.

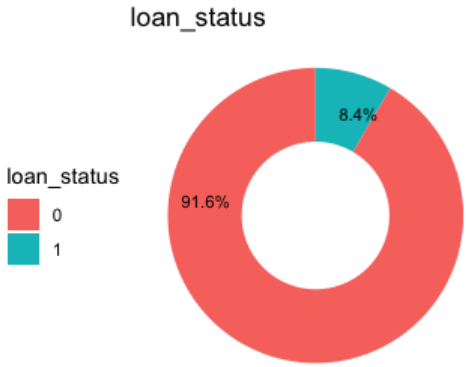
Purpose



This includes different type of the purposes provided by the borrowers during the loan request. Different records include debt_consolidation", "credit_card", "small_business", "major_purchase", "house", "home_improvement", "renewable_energy", "car", "medical", "vacation", "wedding", "educational", "moving" and "other". All these 14 types of recods are then coded from 1 to 14 simultaneously. It is to be noted that, most of the loan are taken for level 1 and level 2 (22.29%) where mode is in level 1 (56.98%).

loan_status

This holds the current status of the loan in the 2nd quarter of 2018. This variable is treated as the dependent variable and converted to binary form for the further analysis. From 9 different levels of data, "Default" and "Charged Off" are coded to 1 (8.44%) and all the other levels are coded to 0 (91.56%).



5 CHAPTER 5: BUILDING PREDICTIVE MODEL (STAGE:2)

This chapter includes the implementation of different potential methods. Based on the different methods, the outcome of the analyses is evaluated. As measure of statistical analysis, correlation and logistics regression have been conducted. Since the main focus of this research includes implementation of fsQCA in financial sector, more detailed explanation and analysis of fsQCA are included in this chapter.

Figure 6 Outline of Chapter 5

5.1 Correlation

Correlation is a quantitative tool that is often used to measure the direction of the relationship between variables and the strength of the change that variables might bring together. The strength and the direction between variables can help to predict the dependent variable based on certain characteristics of the independent variable.

The correlation coefficient is often denoted by “r” and the value ranges between +1 to -1. In this case, if the absolute value is high, it means that the relationship is stronger. Thus, a value of absolute 1 means that, there is a perfect linear relationship. The positive or

negative sign of the coefficient determines the direction of the relationship. The positive sign indicates that the dependent variable changes in the same direction for a change in independent variable. Thus, the graph would be upward. On the other hand, the negative sign indicates that the dependent variable will change in the opposite direction for a change in independent variable. The graph will be downward in this situation.

	term	funded_amnt	int_rate	installment	dti	revol_util	emp_length	purpose	home_ownership	loan_status
term	1	-0.39	-0.39	-0.13	-0.07	-0.07	-0.05	0.08	-0.06	-0.07
funded_amnt	-0.39	1	0.12	0.95	0.05	0.1	0.09	-0.18	0.12	0.02
int_rate	-0.39	0.12	1	0.14	0.15	0.26	-0.01	0.03	-0.05	0.17
installment	-0.13	0.95	0.14	1	0.05	0.12	0.07	-0.17	0.1	0.02
dti	-0.07	0.05	0.15	0.05	1	0.15	0.02	-0.07	0.03	0.04
revol_util	-0.07	0.1	0.26	0.12	0.15	1	0.04	-0.12	-0.02	0.06
emp_length	-0.05	0.09	-0.01	0.07	0.02	0.04	1	-0.01	0.16	-0.01
purpose	0.08	-0.18	0.03	-0.17	-0.07	-0.12	-0.01	1	0.01	-0.01
home_ownership	-0.06	0.12	-0.05	0.1	0.03	-0.02	0.16	0.01	1	-0.03
loan_status	-0.07	0.02	0.17	0.02	0.04	0.06	-0.01	-0.01	-0.03	1

Table 6 Pearson Correlation Coefficient between the variables

Based on the correlation coefficient table and the plot, the Pearson's r value shows that most of the values are close to zero. It shows that there is non-linear relationship between the variables. But, Pearson's r = 0.95 between "funded_amnt" and "installment" shows that there is a strong linear relationship between these variables. Since the time period of the loan repayment is limited to either 36 months or 60 months, the only possible way of distributing monthly payment is through the installment amount. The higher the "funded_amnt" is, the more the loan installment amount. Here, both of these variables are included in the analysis since the relationship with the dependent variable is non-linear and needs to be analyzed. While all the numeric predictor variables show positive

correlation with the dependent variable, categorical variables (e.g. “term”, “emp_length”, “purpose”, “home_ownership”) share negative correlation.

P-value – The analysis is conducted at 95% confidence interval. According to the output, when $\propto = 0.05$, P-value of all the variables is 0. This means that there is no linear relationship between the variables, but the variables are statistically significant since the P-value = 0 (P-value $\leq \propto$). From this correlation coefficient, it is evident that these predictor variables are related to the independent variable.

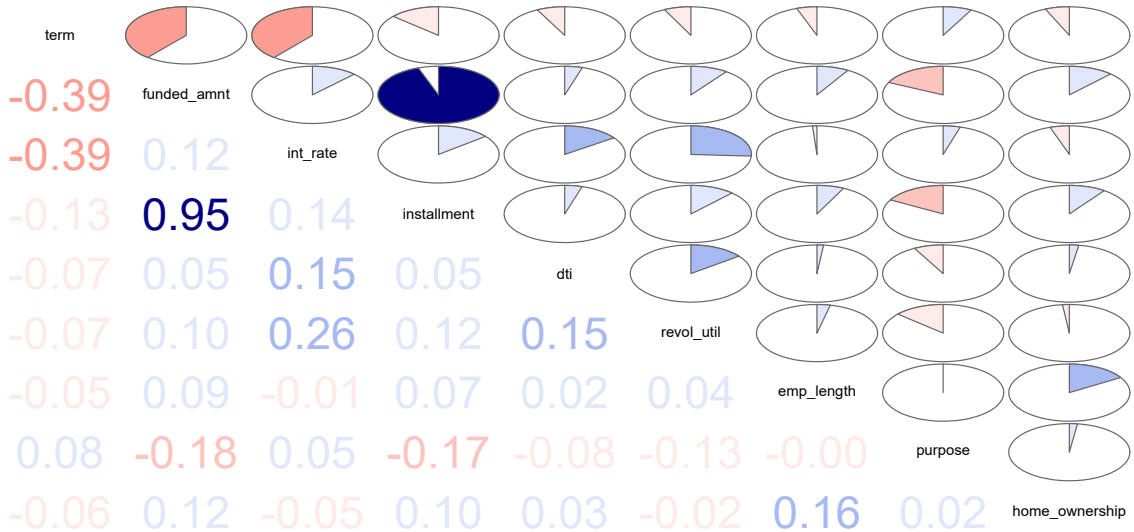


Figure 7 Correlation graph

5.2 Logistic Regression

Logistic regression is often used to predict the class or category of dependent variable based on single or multiple predictor variables. By the class of category, it implies that the dependent variable is supposed to have a categorical variable which can have two possible outcomes. This can be either yes or no, either positive or negative, either 0 or 1. Thus for the purpose of analysis both of the categories are converted to a binary outcome (0 or 1). It means that the analysis uses predictor variables to predict one of the status of the binary outcome. Due to this, different other synonyms e.g. binomial logistic regression, binary logistic regression and logit model are used interchangeably. It is to be noted that, the output from logistic regression does not directly return the class of observations rather it helps to estimate the probability of class membership. So, the

analysis solely depends on the researcher where the cutoff point of the probability is decided based on the type of research.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-5.2952	0.2585	-20.4807	0.0000
term 60 months	0.2972	0.0138	21.6050	0.0000
funded_amnt	0.0000	0.0000	-14.2130	0.0000
int_rate	0.0945	0.0008	118.2632	0.0000
installment	0.0008	0.0001	12.5892	0.0000
dti	0.0032	0.0002	16.0519	0.0000
revol_util	0.0044	0.0001	38.1769	0.0000
emp_length10+ years	-0.0573	0.0083	-6.9413	0.0000
emp_length2-5 years	-0.0188	0.0082	-2.2841	0.0224
emp_length6-9 years	0.1154	0.0091	12.7504	0.0000
purposecredit_card	0.1378	0.0303	4.5417	0.0000
purposedebt_consolidation	0.2647	0.0299	8.8551	0.0000
purposeeducational	0.8473	0.1528	5.5441	0.0000
purposehome_improvement	0.1733	0.0318	5.4465	0.0000
purposehouse	-0.0245	0.0465	-0.5262	0.5988
purposemajor_purchase	0.1213	0.0352	3.4459	0.0006
purposemedical	0.1824	0.0385	4.7386	0.0000
purposemoving	0.1967	0.0420	4.6824	0.0000
purposeother	0.0226	0.0317	0.7124	0.4762
purposerenewable_energy	0.4330	0.0954	4.5372	0.0000
purposesmall_business	0.5936	0.0358	16.5714	0.0000
purposevacation	0.0777	0.0453	1.7130	0.0867
purposewedding	0.4559	0.0723	6.3033	0.0000
home_ownershipMORTGAGE	0.9082	0.2567	3.5386	0.0004
home_ownershipNONE	1.3535	0.5135	2.6357	0.0084
home_ownershipOTHER	1.7474	0.3330	5.2475	0.0000
home_ownershipOWN	1.0055	0.2568	3.9161	0.0001
home_ownershipRENT	1.1707	0.2567	4.5614	0.0000

Table 7 Estimate from the Logistics regression

Statistical significance

Confidence interval of 95% has been followed in this analysis. For most of these variables p-value<0 shows that those variables are statistically significant. Here, "purposehouse", "purposevacation" are not statistically significant since the p-value is higher.

Deviance Residuals

From the Deviance Residuals table of the logistic regression, the outputs are Min: -2.8400; 1Q: 0.2958; Median: 0.3661; 3Q: 0.4461; Max: 2.2249. Since the centered value is close to 0, it means that they are quite symmetrical.

Estimate

The estimate column shows that there are both positive and negative signs for different variables. The positive sign for these predictor variables means that loan status is more likely to be defaulted due to these variables. On the other hand, the negative sign shows that those variables are less likely to impact loan status to be defaulted (those will contribute to non-default).

For several variables, the estimate is not shown for one of the levels, because they are the opposite of the mentioned level. As an example, term 60 months has an estimate of 0.29 which means that the borrowers who choose loan period of 60 months are more likely to default than those who select 36 months.

The estimate of the predictor variables shows that most of the variables has positive sign which shows that presence of these attributes in a borrower might lead to loan default. The high estimate value of "home_ownership", "purposeeducation" indicates that these might have a higher impact on the loan status to be defaulted. Oppositely, "purposehouse", "emp_length10+" years, "emp_length2-5" years has negative estimate which indicates that borrowers with these attributes are less likely to default or more likely to be in non-default category.

Odds Ratio

Odds ratio helps to measure the association between the outcome variable and a predictor variable. Thus, for a one-unit change in the estimate of the predictor variable, the expected

change in log odds is expressed by the odds ratio. Based on the logistic beta coefficients, it measures the ratio of the odds whether an event will occur when the predictor variable is present in the equation (where event = 1; x = 1) compared to the odds of the event occurring in the absence of the predictor variable (x = 0). The following formula means that for a given predictor value (x1), the beta coefficient of that variable (β_1) corresponds to the log of the odds ratio for that predictor.

$$logit(p) = \log(\frac{p}{1-p}) = \beta_0 + \beta_1 * x1$$

As an example, for one of the predictor variables, "int_rate", the regression coefficient is 0.0945 and exp (0.0945) = 1.099058582. This means that one unit increase in the "int_rate" will increase the odds of being "loan_status" to be default by 1.10 times. In other words, there might be 10% increase in the odds of being the loan_status to be default, for a one-unit increase in "int_rate". So, the odds that the event will occur ("loan_status" = 1) is 1.10 times higher when the predictor "int_rate" is present ("int_rate" = 1) versus "int_rate" is absent ("int_rate" = 0). If there are more than one predictor variable in the formula, the odds ratio is calculated keeping the other variables at a fixed value. If "installment" is also added in the formula, keeping "int_rate" at a fixed value, the odds ratio will be calculated from the coefficient.

Variable	**EXP**	**Variable**	**EXP**
term 60 months	1.346050865	purposemajor_purchase	1.128960788
funded_amnt	0.999971118	purposemedical	1.200084149
int_rate	1.099058582	purposemoving	1.217382901
installment	1.000808140	purposeother	1.022827754
dti	1.003218197	purposerenewable_energy	1.541905013
revol_util	1.004425146	purposesmall_business	1.810579437
emp_length10+ years	0.944302550	purposevacation	1.080752030
emp_length2-5 years	0.981407430	purposewedding	1.577625503
emp_length6-9 years	1.122344498	home_ownershipMORTGAGE	2.479888274

purposecredit_card	1.147772285	home_ownershipNONE	3.870876281
purposedebt_consolidation	1.303081253	home_ownershipOTHER	5.739914540
purposeeducational	2.333229158	home_ownershipOWN	2.733239722
purposehome_improvement	1.189281603	home_ownershipRENT	3.224272217
purposehouse	0.975821203		

Table 8 Odds ratio of the predictor variables

The values of the odds ratio measure the association between an exposure and an outcome. If the value is greater than 1, it means the exposure is associated with higher odds of outcome. If the value is less than 1, it means the exposure is associated with lower odds of outcome. Finally, if the exposure equals 1, it means that the exposure does not affect odds of outcome. From the odds ratio table, it is clear that "home_ownership" has higher odds ratio followed by "purpose". The other predictor variables have odds ratio equivalent to 1 and in some cases less than 1. It means that the effect of the odds ratio of those variables to "loan_status" will be lower.

Wald test

Chi-squared test, $X^2 = 20.8,\ df = 1,\ P(> X2) = 5.1e-06 \approx 0.013$. So, the null hypothesis can be rejected. Thus, it shows that predictor variables are statistically significant and might lead to better prediction. Based on the Chi-square, those variables can be included into the model.

Based on all of these statistical measures, a model can be presented which can effectively predict the "loan_status" whether it will result in "Default" or "Non-default".

$$logit(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_i * x_i$$

$$logit(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 * \mathbf{term} + \beta_2 * \mathbf{funded\_amnt} + \beta_3 * \mathbf{int\_rate} + \beta_4 * \mathbf{installment} + \beta_5 * \mathbf{dti} + \beta_6 * \mathbf{revol\_util} + \beta_7 * \mathbf{emp\_length} + \beta_8 * \mathbf{purpose} + \beta_9 * \mathbf{home\_ownership}$$

Prediction accuracy

Based on the confusion matrix different accuracy measures for the prediction can be calculated. Among all these, the most important measures can be sensitivity and specificity. Sensitivity- a proportion of actual positives identified correctly; specificity- a proportion of actual negatives identified correctly. Although there is a tradeoff between the sensitivity and specificity, selecting a perfect threshold is a challenge. Here, when the threshold is greater than 0.2, the model performs better compared to 0.5 and 0.7.

	Training Set (>0.7)	**Test Set (>0.7)**	**Training Set (>0.5)**	**Test Set (>0.5)**	**Training Set (>0.2)**	**Test Set (>0.2)**
Accuracy	0.91555	0.91555	0.91554	0.91552	0.91554	0.91552
Sensitivity	0.00000	0.00000	0.00003	0.00003	0.00003	0.00003
Specificity	0.99999	0.99999	0.99997	0.99995	0.99997	0.99995
Precision	0.00000	0.00000	0.07500	0.04545	0.07500	0.04545
Negative predictive value	0.91556	0.91556	0.91556	0.91556	0.91556	0.91556

Table 9 Accuracy matrix for model prediction

ROC Test

From the ROC test, the area under the curve, AUC = 0.6868, shows that the model can discriminate between positive and negative classes better than a random model. Although a score of 1.0 represents a perfect model, but 0.6838 is considerably a good score for the prediction.

5.3 fsQCA Analysis

The method fsQCA uses combinational logic, attributes from fuzzy set theory and Boolean minimization to analyze what combinations are necessary to produce an outcome. The basic process flow includes multiple steps. Firstly, from the available variables, the most important variables will be selected and categorized. Secondly, fuzzification of those input and output variables will be conducted based on the membership of the degree of certainty. Thirdly, inference rules will be applied. Fourthly, the output from the independent variables will be produced and finally the result will be analyzed for developing the model. Similar to this, Bennouna & Tkiouat (2018) have followed a 5-step process in their research using Matlab to process the data. In this

research a software called “R programming” is used instead of Matlab. From R (version 3.5.2), a package titled “QCA” is used to process the data for fsQCA.

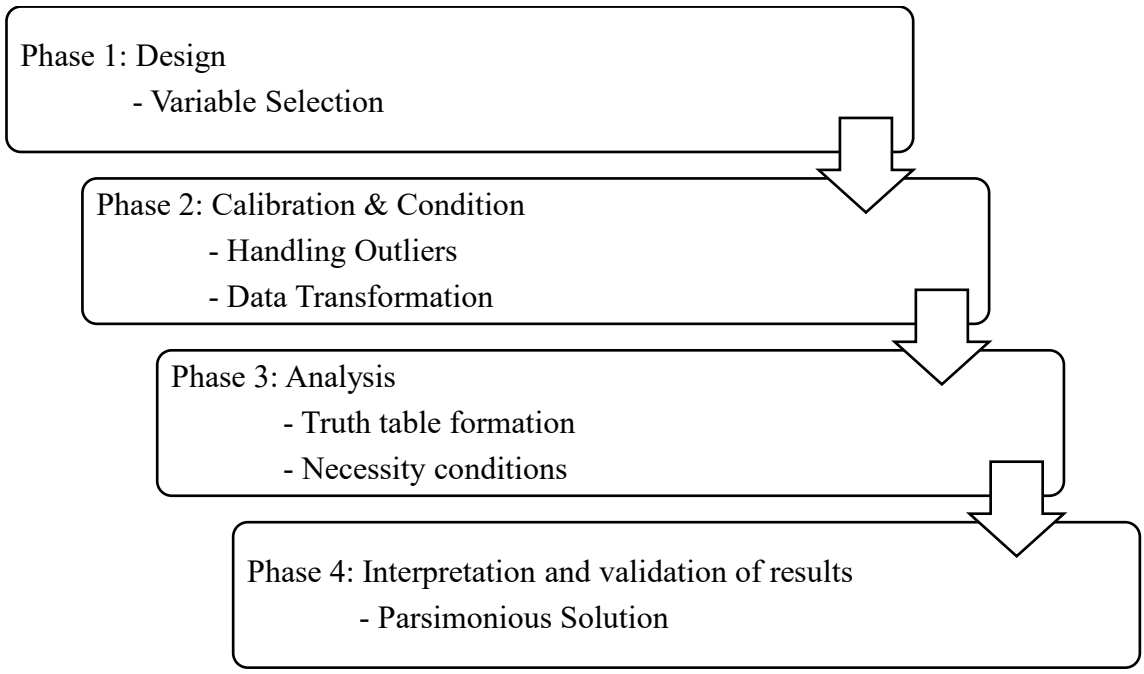


Figure 8 Fuzzy Set QCA analysis process

5.3.1 Phase 1: Design

Variable selection

From the raw dataset having 2004091 x 151 values, an initial list of predictor variables was selected to run logistics regression. Based on the correlation table and the graph those available variables were filtered again to choose optimal number of variables for fsQCA since larger number of predictor variable might make the membership function very complex to explain. Selected predictor variables are "int_rate","dti", "revol_util","emp_length" and "home_ownership". The independent variable is "loan_status" which had value of 0 or 1 (where 1 means loan default and 0 means non-default).

5.3.2 Phase 2: Calibration and Condition

Handling outliers

Because big data includes suffers from because of outliers. The numerical variables “int_rate”, “dti”, “revol_util” all had outliers. The summary of the dataset showed that,

int_rate has value ranging from 5.31 to 30.99 with a mean value of 13.11. Also, from the density graph of "int_rate" (showed above), it is clearly visible that the density after 20 is very minimum. Thus, "int_rate" variable having value greater than 20 is mutated to 20. Similarly, the values greater than 40 in "dti" are mutated to 40 and the values greater than 60 are mutated to 60.

Data transformation

As an initial requirement of fsQCA to proceed, all the variables must be within a range of 0 to 1. Here, 0 means non-membership and 1 means full membership. A fuzzy score of 0.5 means neither in nor out of the set. This score brings the maximum ambiguity.

All the numeric variables are converted simply by deducting the variables' minimum value from the value itself and later dividing that value with the difference from maximum and the minimum value. The formula used for the transformation is –

$$\frac{x - \text{minimum}(x)}{\text{maximum}(x) - \text{minimum}(x)}$$

For the categorical variables, those variables had many levels which were considerably transformed to value between 0 and 1. Here, variable "emp_length" can hold values of four different levels. And the variable "home_ownership" can hold values of six different levels. Both of these are later transformed by the following ways.

emp_length	emp_length less than 1 year = 0.05 emp_length between 1 and 2 year= 0.33, emp_length between 2 and 3 year and equal to 3 year = 0.66, emp_length greater than 3 year = 0.95,
home_ownership	home_ownership less than 2 = 0.05 home_ownership between 2 and 3 = 0.25 home_ownership between 3 and 4 = 0.45 home_ownership between 4 and 5 = 0.65 home_ownership between 5 and 6 = 0.85 home_ownership equal to 6 or above = 0.95

5.3.3 Phase 3: Analysis

In order to understand the scenario where the fsQCA performs its best the analysis is divided into three different scenario where different size of sample and different values of the tools are used.

Scenario	n size	incl.cut	n.cut	Prediciton
Scenario 1	500000	0.92	5	Non-default
Scenario 2	500000	0.095	5	Default
Scenario 2	100000	0.50	5	Default

In every scenario, conditions for the analysis are "int_rate","dti", "revol_util", "emp_length", "home_ownership" and the outcome is “loan_status”.

Truth table formation

In this stage, as a “outcome” variable “loan_status” is selected which has value of either 0 or 1. Basically, truth table helps to explain the outcome based on at least two membership conditions which might be necessary for the outcome to occur. This table consider each case as combination of the characteristics selected and the cases with exact same configuration are considered to be the same type of case. Generally, the number of combinations in a truth table are 2k where k is the number of casual conditions mentioned by the researchers. In our case, based on both numerical and categorical variables which are transformed to numerical dataset, QCA package is used in R environment. All the predictor variables "int_rate","dti", "revol_util","emp_length", "home_ownership" are considered as conditions in the truth table formation.

	INT_RATE	DTI	REVOL_UTIL	EMP_LENGTH	HOME_OWNERSHIP	OUT	n	incl	PRI
3	0	0	0	1	0	1	26607	0.947	0.947
4	0	0	0	1	1	1	6	0.945	0.945
12	0	1	0	1	1	1	6	0.943	0.943
1	0	0	0	0	0	1	25127	0.943	0.943
11	0	1	0	1	0	1	10966	0.943	0.943
2	0	0	0	0	1	1	13	0.941	0.941
9	0	1	0	0	0	1	9375	0.939	0.939
7	0	0	1	1	0	1	57575	0.935	0.935
8	0	0	1	1	1	1	16	0.934	0.934
16	0	1	1	1	1	1	21	0.933	0.933
5	0	0	1	0	0	1	50519	0.932	0.932
15	0	1	1	1	0	1	38757	0.931	0.931
6	0	0	1	0	1	1	23	0.930	0.930
14	0	1	1	0	1	1	15	0.929	0.929
13	0	1	1	0	0	1	30786	0.928	0.928
28	1	1	0	1	1	1	6	0.925	0.925
18	1	0	0	0	1	1	7	0.923	0.923
19	1	0	0	1	0	0	10550	0.919	0.919
27	1	1	0	1	0	0	6362	0.917	0.917
17	1	0	0	0	0	0	12041	0.916	0.916
25	1	1	0	0	0	0	6714	0.914	0.914
24	1	0	1	1	1	0	18	0.914	0.914
32	1	1	1	1	1	0	20	0.913	0.913
22	1	0	1	0	1	0	35	0.911	0.911
30	1	1	1	0	1	0	27	0.910	0.910
23	1	0	1	1	0	0	54583	0.905	0.905
21	1	0	1	0	0	0	49753	0.902	0.902
31	1	1	1	1	0	0	60012	0.899	0.899
29	1	1	1	0	0	0	49105	0.897	0.897

Figure 9 Truth table for Scenario 1

	INT_RATE	DTI	REVOL_UTIL	EMP_LENGTH	HOME_OWNERSHIP	OUT	n	incl	PRI
29	1	1	1	0	0	1	49105	0.103	0.103
31	1	1	1	1	0	1	60012	0.101	0.101
21	1	0	1	0	0	1	49753	0.098	0.098
23	1	0	1	1	0	1	54583	0.095	0.095
30	1	1	1	0	1	0	27	0.090	0.090
22	1	0	1	0	1	0	35	0.089	0.089
32	1	1	1	1	1	0	20	0.087	0.087
24	1	0	1	1	1	0	18	0.086	0.086
25	1	1	0	0	0	0	6714	0.086	0.086
17	1	0	0	0	0	0	12041	0.084	0.084
27	1	1	0	1	0	0	6362	0.083	0.083
19	1	0	0	1	0	0	10550	0.081	0.081
18	1	0	0	0	1	0	7	0.077	0.077
28	1	1	0	1	1	0	6	0.075	0.075
13	0	1	1	0	0	0	30786	0.072	0.072
14	0	1	1	0	1	0	15	0.071	0.071
6	0	0	1	0	1	0	23	0.070	0.070
15	0	1	1	1	0	0	38757	0.069	0.069
5	0	0	1	0	0	0	50519	0.068	0.068
16	0	1	1	1	1	0	21	0.067	0.067
8	0	0	1	1	1	0	16	0.066	0.066
7	0	0	1	1	0	0	57575	0.065	0.065
9	0	1	0	0	0	0	9375	0.061	0.061
2	0	0	0	0	1	0	13	0.059	0.059
11	0	1	0	1	0	0	10966	0.057	0.057
1	0	0	0	0	0	0	25127	0.057	0.057
12	0	1	0	1	1	0	6	0.057	0.057
4	0	0	0	1	1	0	6	0.055	0.055
3	0	0	0	1	0	0	26607	0.053	0.053

Figure 10 Truth table for Scenario 2

	INT_RATE	DTI	REVOL_UTIL	EMP_LENGTH	HOME_OWNERSHIP	OUT	n	incl	PRI
29	1	1	1	0	0	1	12156	0.554	0.554
31	1	1	1	1	0	1	15052	0.550	0.550
21	1	0	1	0	0	1	12286	0.542	0.542
23	1	0	1	1	0	1	13678	0.535	0.535
30	1	1	1	0	1	1	5	0.517	0.517
22	1	0	1	0	1	1	6	0.515	0.515
25	1	1	0	0	0	1	1485	0.501	0.501
17	1	0	0	0	0	0	2762	0.499	0.499
27	1	1	0	1	0	0	1450	0.493	0.493
19	1	0	0	1	0	0	2366	0.488	0.488
13	0	1	1	0	0	0	4799	0.458	0.458
15	0	1	1	1	0	0	6191	0.446	0.446
5	0	0	1	0	0	0	8286	0.441	0.441
7	0	0	1	1	0	0	8968	0.430	0.430
9	0	1	0	0	0	0	1399	0.411	0.411
11	0	1	0	1	0	0	1514	0.398	0.398
1	0	0	0	0	0	0	3682	0.392	0.392
3	0	0	0	1	0	0	3716	0.379	0.379

Figure 11 Truth table for Scenario 3

For all these truth tables in different scenarios, each of the rows shows the possible combinations that a configuration can include and the output of that configuration. The output is explained by the column "OUT" and the other columns represents the conditions. "inclS" and "PRI" represents the Proof of Fit of the measure.

As a general rule of fsQCA, the variables do not contain a value equal to 0.5 in the membership function. If the calibrated value of those variable is higher than 0.5 the truth table shows 1 and 0 otherwise. As a measure of truth table minimization, inclusion cut-off can be implemented carefully thus for different scenario different cut-off values have been used. For scenario 1, 2 and 3, cut-off score of 0.92, 0.095 and 0.50 have been used consecutively. As a result, the truth tables in different scenarios, shows considerable amount of casual conditions where "OUT" is 1. It is also to be noted that none of the truth tables show "?" which means that there is no counterfactual (an unobserved configuration having logical remainder).

Necessity conditions

Using this tool from fsQCA, researches can understand the combination which can be considered significant. It shows inclN, RoN and covN as the measures. These different measures have different benchmark which normally researchers follow. For the inclusion 0.85 is considered a benchmark as suggested by Ragin (2008). For RoN and covN, it is considered to be 0.6 as suggested by Dusa (2017).

covN

covN is a measure of trivial configuration. Necessity of trivial configuration is used to determine the relevance of conditions for the outcome. covN explains the proportion of the necessary conditions that is covered by the outcome. Th formula used to calculate covN

$$covN_{X \Leftarrow Y} = \frac{\sum \min(X, Y)}{\sum X}$$

RoN

RoN is also a measure of trivial configuration. It measures the relevance of the solution. If the value is equal to 1 in most of the cases the explanation might be meaningless for the analysis.

		inclN	RoN	covN
1	REVOL_UTIL+EMP_LENGTH+home_ownership	0.930	0.464	0.914
2	int_rate+dti+REVOL_UTIL+home_ownership	0.924	0.489	0.915
3	int_rate+DTI+REVOL_UTIL+home_ownership	0.922	0.495	0.915
4	int_rate+REVOL_UTIL+emp_length+home_ownership	0.926	0.479	0.915
5	INT_RATE+dti+REVOL_UTIL+emp_length+home_ownership	0.925	0.476	0.913
6	INT_RATE+DTI+REVOL_UTIL+emp_length+home_ownership	0.922	0.486	0.913

Figure 12 Necessity measures for Scenario 1

		inclN	RoN	covN
1	INT_RATE	0.676	0.625	0.577
2	dti	0.504	0.636	0.480
3	REVOL_UTIL	0.807	0.375	0.520
4	EMP_LENGTH	0.556	0.608	0.495
5	home_ownership	0.822	0.314	0.505
6	INT_RATE*REVOL_UTIL	0.601	0.681	0.574
7	INT_RATE*home_ownership	0.622	0.660	0.570
8	REVOL_UTIL*home_ownership	0.704	0.498	0.519
9	EMP_LENGTH*home_ownership	0.504	0.663	0.500
10	INT_RATE*REVOL_UTIL*home_ownership	0.565	0.701	0.568
11	int_rate+DTI	0.571	0.554	0.472
12	int_rate+emp_length	0.560	0.547	0.462
13	DTI+revol_util	0.549	0.627	0.504
14	DTI+emp_length	0.641	0.547	0.510
15	DTI+HOME_OWNERSHIP	0.508	0.684	0.518
16	revol_util+emp_length	0.509	0.644	0.489

Figure 13 Necessity measures for Scenario 3

The scores in scenario 1 and scenario 3 shows that there are number of combinations where the value is more than 0.85. The scores in Scenario 2 is included in appendix which also responds according to the cut-off value. Therefore, those combinations can be

potentially significant which can determine the loan status. It is to be noted that in Scenario 1, positive loan status directs to non-default since neg.out = TRUE is used for the analysis. The other two cases Scenario 2 and 3 used neg.out = FALSE which means that the outcome to be predicted is loan default.

For fsQCA the uppercase style of the conditions means close to 1 and the lowercase translates to 0 since all the values were transformed to a range of 0 to 1 for the analysis.

5.3.4 Phase 4: Interpretation and validation of results

inclS

One of the most important proof of fit of QCA is fuzzy sufficiency inclusion or consistency. This is in effect when both the conditions and the outcome in all the cases exists when the conditions occur.

$$inclS_{X \Rightarrow Y} = \frac{X \cap Y}{X}$$

PRI (Proportional Reduction in Consistency)

It measures the phenomenon of simultaneousness subset relations which is considered the best option for fsQCA. The calculation of PRI subset relations is based on the conditions where X seems to be sufficient of Y and the negation of Y denotes which are logically not possible.

$$PRI = \frac{\sum \min(X, Y) - \sum \min(X, Y, \sim Y)}{\sum X - \sum \min(X, Y, \sim Y)}$$

According to Ragin (2015), PRI score smaller than the inclusion score the tested conditions are sufficient for Y and not also for the negation. The smaller PRI score also mean that the score above 0.6.

Raw coverage (covS)

Raw coverage is a measure of sufficiency. It shows how much of the outcome is explained by the conditions. It considers the summation of the min values of both X and Y and divides by the summation of Y.

$$covS = \frac{\sum \min(Y, X)}{\sum Y}$$

Unique coverage (covU)

Unique coverage explains how much of the explanation can be uniquely attributed to that set. It is the unique coverage which excludes the other coverages or solely explains as the difference between the raw coverage and other overlaps.

Parsimonious Solution results: Scenario 1

```
n OUT = 1/0/C: 249825/249220/0
  Total      : 499045

M1: int_rate + revol_util*HOME_OWNERSHIP => loan_status

                              inclS  PRI    covS   covU
-----------------------------------------------------------
1  int_rate                   0.944  0.944  0.504  0.414
2  revol_util*HOME_OWNERSHIP  0.935  0.935  0.099  0.009
-----------------------------------------------------------
   M1                         0.942  0.942  0.513
```

Figure 14 Parsimonious solution results with PoF (Scenario 1)

From the results of parsimonious solutions (Scenario 1) the measures of PoF, inclS>0.8 and PRI>0.6. Although covS is not below 0.6 but close to the benchmark. Therefore "int_rate" can be considered potential which can produce the outcome. Here, "int_rate" means that the lower the amount of the interest rate the higher the chance is that the loan will not default. The second expression also consists of a high value which explains that the combination "revol_util" and "HOME_OWNERSHIP". For a specific borrower, if the usage of credit out of all available revolving credit is low and the type of home ownership is anything excluding "Rent" or "Mortgage", there is less chance that the loan will default.

Here, finally, in Scenario 1, it says that either low interest rate or a combination of lower revolving credit usage and ownership of home can be potentially significant to determine the outcome of loan status. In this case, neg.out = TRUE which determines 0 or in other word non-default.

Parsimonious Solution results: Scenario 2

```
n OUT = 1/0/C: 213453/285592/0
  Total      : 499045

M1: INT_RATE*REVOL_UTIL*home_ownership => LOAN_STATUS

                                        inclS  PRI    covS   covU
-------------------------------------------------------------------
1  INT_RATE*REVOL_UTIL*home_ownership  0.108  0.108  0.563    -
-------------------------------------------------------------------
   M1                                  0.108  0.108  0.563
```

Figure 15 Parsimonious solution results with PoF (Scenario 2)

In scenario 2, the target is to see if it can predict the loan status of the borrowers which tends to default. An inclusion cut-off value of 0.095 is used. One important factor to remember is that the sample dataset that is used holds a proportion of 0.92:0.08. Since the data is too much distributed and the occurrence of the outcome has a very long range the value received in the PoF might be used considering this case as an exception. The result shows that the interest rate AND the higher the usage of the revolving credit AND the situation where home is rented or mortgaged, the more there is chance that the loan will be defaulted considering these conditions.

Parsimonious Solution results: Scenario 3

```
n OUT = 1/0/C: 54668/45133/0
  Total      : 99801

M1: INT_RATE*REVOL_UTIL + INT_RATE*DTI*emp_length => LOAN_STATUS

                           inclS  PRI    covS   covU
-----------------------------------------------------
1  INT_RATE*REVOL_UTIL     0.574  0.574  0.601  0.337
2  INT_RATE*DTI*emp_length 0.555  0.555  0.276  0.011
-----------------------------------------------------
   M1                      0.573  0.573  0.612
```

Figure 16 Parsimonious solution results with PoF (Scenario 3)

In the case of Scenario 2, it is visible that the result is conditionally biased due to the higher range of the non-default to default ratio in the outcome variable. In order to see if the same result persists, same conditions and measures are used but with a different

inclusion cut-off value (incl.cut = 0.50) from a sample dataset (n=100000) where the proportion stands to 1:1. In order to maintain the transparency, out of the raw dataset two separate dataset of non-default and default values were created. Later, for both of these two different datasets 50000 records were selected and then merged together. Finally, the merged dataset was rearranged and used for the analysis for Scenario 3. Although the PoF is still below the traditional benchmarks but compared to the proportion of the dataset used and the incl.cut value the result shows better output quality with an increase in the quality of the prediction. The M1 predicts two different expressions that might lead to the loan status being defaulted.

Expression 1 explains that the high interest rate AND high usage of the revolving credit are potential conditions that can lead loan status to be defaulted.

Expression 2 means that high interest rate AND higher total debt payments from the borrower's income AND employment length below 5 years might be potential contributor towards loan default.

6 CHAPTER 6: DISCUSSION

For traditional methods like logistics regression it does not necessitate the coding of the dependent variable in the exact order. The coefficient will result same no matter how the variable is coded either from 0 – 1 or 1 – 0. But, the case of fsQCA is different, since the casual relationships in fsQCA are assumed to be symmetrical. In this research, the occurrence of default is not just the inverse or logical negation. The analysis conducted for the default does not explain the non-default if the outcome is inversed. It means that there has to be separate analysis for both of the output. Here, if a businessman wants to measure the riskiness of the loan for a specific borrower, that businessman should not just depend on the sole result derived from the analysis done for predicting default rather he/she should analyze the expressions for the non-default also. Based on the subjective expertise, the businessman should choose the balance between these two if the loan should be sanctioned or if it needs the rate of returned to be increased to balance the risk. Unlike logistics regression, fsQCA is providing more details about the conditions that might potentially lead to the same outcome. Based on the membership functions, it gives more customization about the decision-making process. The lender can decide which condition of the predictor variable is significantly responsible for the outcome. In this study, the logistics regression model included "term", "funded_amnt", "int_rate", "installment", "dti", "revol_util", "emp_length", "purpose" and "home_ownership" which lead to the occurrence of the outcome. These variables do not include any range up to which can lead to the outcome. For example, it does not say if the high interest rate or low interest rate tend to affect the loan status more. Here, if we see the results of fsQCA from Scenario 3 it says that high interest rate and high revolving utility usage might be potential to loan default OR high interest rate high debt-to-income ratio and less employment length can be significant for the occurrence of loan default. Comparing these two explanations, it can be said that the explanation of fsQCA with the membership functions is more specific rather than ambiguous. This is more practical since fsQCA assigns full membership, partial membership and no membership to the conditions. This feature of fsQCA makes it more interpretable and usable in real life situation where results might not always be "yes" or "no" but a range between yes or no. The option of producing alternative expressions in fsQCA make it adaptable in the real life. The result is more detailed compared to the logistics regression. The expressions with conditions are

identifiable with the membership. This feature makes the output more detailed and specific. Practically, all of the information about the borrowers might not be available in real life. Therefore, the alternative models might be a tool that can serve the purpose of both the parties.

For the purpose of processing big data, fsQCA might perform better since there has been many issues for the logistics regression model with larger n-sized datasets. As we know that the casual relationship of the logistics regression model is negatively evidenced if the dependent variable occurs, but the independent variable does not. This means that the factor which has an impact on the dependent variable in a subset become invisible in the regression analysis. Contrarily the fsQCA identifies the patterns across different subsets of cases. As a result, fsQCA can handle larger dataset without severe data requirements.

Limitations in fsQCA

As we know fsQCA need the calibration of the data, other methods (e.g. Multiple Regression Analysis) does not require it. A researcher needs to calibrate the conditions before running the analysis. Sometimes defining the membership function can be a challenge without proper knowledge of the research. This membership function leads to another limitation of fsQCA - the knowledge base of the experts. Fuzzy inference system can only be used where membership functions are known in advance and the rule structure is essentially predetermined by the experts. (Malhotra & Malhotra, 2002). This necessitates prior knowledge of the experts and might hamper transparency (Vis, 2012). Finally, unlike traditional statistical tools, fsQCA does not estimate the net effect of a single variable on the outcome but focuses on the combination of the plausible outcomes. If a researcher wants to analyze the net effect of any independent variable on the dependent variable, those traditional methods work best (Vis, 2012). In this research, it is observed that the QCA package in R cannot handle larger data size which exceeds approximately 500000 records. Thus, the same technique can be implemented using different package or software to investigate the consistency of the outcome. Comparison between logistic regression and fsQCA are also done in narrowly due to time restriction. Further study be conducted using the complex and intermediate solution from fsQCA to compare with logistic regression or other traditional tools. Since this research completely focuses on the P2P platform, in real life, it is also important to check if fsQCA performs consistently in other financial sectors also.

7 CHAPTER 7: CONCLUSION

The primary objective of this research is to understand the key differences that the fsQCA can offer compared to the logistic regression. One of the important findings from the odds ratio of the logistic regression is that the logit model includes 9 different variables. Almost all of the variables share higher exposure leading to higher odds to the outcome. Although logistic regression included many variables, fsQCA (with a smaller number of variables) showed a potential increment in the performance compared to the proportion of the default to non-default rate. Since many researchers emphasized to use the standard benchmarks of PoF for fsQCA, it is important that the researcher determines the threshold based on the type of research and the volume of the data. In scenario 1, fsQCA performed best to identify the core conditions leading towards the occurrence of the outcome. It showed non sharp boundaries of the conditions towards predicting non-default instead of default. On the other hand, in scenario 2 and 3, fsQCA performed efficiently to identify the potential conditions that might lead to loan default. If the results are then compared to the traditional logistic regression, it is noticeable that fsQCA narrowed down the conditions but included non-sharp boundaries and alternative configurations for the decision makers. This allows more scope towards decision making process. Fuzzy set QCA (fsQCA) also predicts the outcome better than the percentage of the original dataset. Based on these results, this research finds out the RQ 1 to be correct and worth acceptable instead of logistic regression.

The logit model shows the association of these predictor variables to the outcome having a sharp boundary. But, in real life, the decision makers might be interested into the values inside the boundary rather than straight “yes” or “no”. It is to be noted that, Felício et al., (2016) mentioned in their research that many organizations might require different solutions that lead to same result. Here, fsQCA offers more customized outcome by showing different combinations of the conditions in different configurations leading to the same outcome. Boratyńska & Grzegorzewska (2018) also emphasized fsQCA over the logistic regression. Agreeing with these statements, the parsimonious solution of fsQCA also suggest that there are different configurations showing membership functions in the output. As Chari et al., (2016) mentioned, these alternative configurations lead more options for the businesses towards selecting the optimal trade-off. Opposite to the logistics regression, this research showed the applicability of fsQCA in financial sector

where different casual combinations of casual conditions is demanded. Since the objective of any organization might differ due to the goal, firm size and overall strategy, the overall importance of the predictor variables to achieve the output might not be enough in all cases (as conducted in logistics regression). In real life, organizations always seek to minimize the risk based on the available alternatives. They might want to find out the interplay between different conditions and the nature of casual patterns. This research shows that there are several expressions from fsQCA results which can supplement the findings compared to the logistics regression. This additional insight into the outcome can assist the businesses to cope up with the real-life complex scenarios. Since, fsQCA offers alternative solutions, businesses can determine the optimal solution based on their need, real life complexity and customer need. It showed that there is a key difference that fsQCA can make when decision makers need to handle casually complex problems in real life. Since the usage of fsQCA in business sector is comparatively low, it shows the prospects of using fsQCA in business sectors for the decision makers. Thus, it can be said that, fsQCA performed efficiently in this analysis. Thus, it is suggested for the financial industry for the optimization of the decision-making process.

Since the original dataset included almost 2 million records, due to the limitation of R software, a sample of 500000 is used for fsQCA. In order to overcome any biasness, random sampling is used several times. On the other hand, logistic regression could process the whole dataset. In this case the output performance of fsQCA is slightly better than the logistic regression. It seemed that the QCA package of R has a limitation of processing larger volume of data rather than fsQCA itself. This research demands further study which can focus on larger “n size” to investigate if fsQCA still performs consistently. Other than the limitation of larger volume, it is worth mentionable that fsQCA handled asymmetric and non-linear data efficiently. In order to handle the outliers, dataset was easily transformed which still produced an acceptable output compared to the logistic regression. Finally, it is to be concluded that this study could not fully answer RQ 4 due to the limitation if R package.

Finally, it is worth mentioning that the outcome from fsQCA does not share the property of reversal. In many cases, the outcome of many traditional tools can be logically reversed to investigate the opposite scenario of the outcome. But, fsQCA does not perform similar to this. If we compare the results from Scenario 1 to Scenario 2 or 3, it is noticeable that the negation of the output does not produce any meaningful result. Rather, the analysis

has to be conducted separately to investigate the conditions of non-default compared to default. This research adds to the existing literatures that fsQCA produce separate outcome for different scenario and thus should not be negated to take decision of the opposite scenario.

Since, software R (version 3.5.2,) is used with the QCA package for analyzing fsQCA, it is shown that this package can handle big data. The only limitation was observed when n size exceeds 500000. The larger n size than this showed maximum memory capacity usage problem and stopped R to respond. Thus, newer versions of software of newer package might be tested in the future research. As a part of further research, fsQCA can be implemented using other packages which can handle n size larger than 500000. Since the number of casual conditions creates more membership functions, an efficient way of handling larger number of casual conditions can also be investigated.

REFERENCES

Özari, Ç., & Ulusoy, V. (2017). Estimation of bankruptcy probabilities by using Fuzzy logic and Merton model: An application on USA companies. Business & Management Studies: An International Journal, 5(4), 211-234.

Abdullayev, T., Umarova, N., Jamalov, Z., & Alekperov, A. (2016). Fuzzy logic and lexicological support of the creation of terminological dictionary of intermediate language in bilingual education. 12th International Conference on Application of Fuzzy Systems and Soft Computing. 102, pp. 390 – 397. Vienna, Austria: Procedia Computer Science.

Abdulrahman, U., Panford, J., & Hayfron-Acquah, J. (2014, May). Fuzzy Logic Approach to Credit Scoring for Micro Finances in Ghana. International Journal of Computer Applications, 94(8).

Allen, M., & Allen, M. (2015). Industrial Relations, and Economic Growth: A Comparative Analysis of the States of South Eastern Europe. Research in International Business and Finance, 33, 167–177.

Apetrei, A., Paniagua, J., & Sapena, J. (2016). Do Financial Crises Moderate Entrepreneurial Recipes? A Comparative Fuzzy Analysis. Journal of Promotion Management, 22(4), 482-495.

Bagherpour, A. (2018, October 15). Official page of University of California. Retrieved from http://economics.ucr.edu/job_candidates/Bagherpour-Paper.pdf

Bajpai, P. (2016, September 27). Retrieved from NASDAQ: https://www.nasdaq.com/article/the-rise-of-peertopeer-p2p-lending-cm685513

Baku, E., & Smith, M. (2010, March 31). Loan delinquency in community lending organizations: Case studies of neighborworks organizations. Housing Policy Debate, 9(1), 151-175. doi:10.1080/10511482.1998.9521289

Bennouna, G., & Tkiouat, M. (2018). Fuzzy logic approach applied to credit scoring for microfinance in Morocco. Procedia Computer Science, (pp. 274-283).

Beynon, M., Jones, P., & Pickernell, D. (2016). Country-level investigation of innovation investment in manufacturing: Paired fsQCA of two models. Journal of Business Research, 69, 5401–5407.

Bhimani, A., Gulamhussen, M. A., & Lopes, S. D. (2013). The Role of Financial, Macroeconomic, and Non-financial Information in Bank Loan Default Timing Prediction. European Accounting Review, 22(4), 739–763. doi:10.1080/09638180.2013.770967

Boratyńska, K., & Grzegorzewska, E. (2018). Bankruptcy prediction in the agribusiness sector: Lessons from quantitative and qualitative approaches. Journal of Business Research, 89, 175-181.

Bourke, P., & Shanmugan, B. (1990). The Management of Financial Institutions. (P. Bourke, & B. Shanmugam, Eds.) Indiana, USA: Addison-Wesley.

Bruett, T., Alternative Credit Technologies, LLC, Echange LLC, & Enterprise Solutions Global Consulting. (2004). Four Risks That Must Be Managed by Microfinance Institutions. Retrieved October 19, 2018, from http://www.arabic.microfinancegateway.org/sites/default/files/mfg-en-paper-four-risks-that-must-be-managed-by-microfinance-institutions-nov-2004.pdf

Chari, S., Tarkiainen, A., & Salojärvi, H. (2016). Alternative pathways to utilizing customer knowledge: A fuzzy-set qualitative comparative analysis. Journal of Business Research, 69, 5494–5499.

Chaudhari, S., & Patil, M. (2014, 12 17). Comparative Analysis of Fuzzy Inference Systems for Air Conditioner. International Journal of Advanced Computer Research, 4(17).

Chen, L.-H., & Chiou, T.-W. (1999). A fuzzy credit-rating approach for commercial loans: a Taiwan case. Omega International Journal of Management Science, 27, 407-419.

Coduras, A., Clemente, J., & Ruiz, J. (2016). A novel application of fuzzy-set qualitative comparative analysis to GEM data. Journal of Business Research, 69, 1265–1270.

Córdova, J. F., Molina, E. C., & López, P. N. (2017, December). Fuzzy logic and financial risk. A proposed classification of financial risk to the cooperative sector. Contaduría y Administración, 62, 1687–1703.

Elliott, T. (2013). Fuzzy set qualitative comparative analysis. UCI. Research Notes: Statistics Group.

Feldman, K., & Treleaven, P. (1994). Intelligent Systems in Finance. Applied Mathematical Finance, 195-205.

Felício, J. A., Duarte, M., & Rodrigues, R. (2016). Global mindset and SME internationalization: A fuzzy-set QCA approach. Journal of Business Research, 69, 1372–1378.

Felício, J., Rodrigues, R., & Samagaio, A. (2016). Corporate Governance and the Performance of Commercial Banks: A Fuzzy-Set QCA Approach. Journal of Small Business Strategy, 26(1).

Fiss, P. (2011, April). Building Better Causal Theories: A Fuzzy Set Approach to Typologies Inorganization Research. The Academy of Management Journal, 54(2), 393-420.

Guo, Y., Zhou, W., Luo, C., Liu, C., & Xiong, H. (2016). Instance-based credit risk assessment for investment decisions in P2P lending. European Journal of Operational Research, 249, 417-426.

Hsiao, Y.-H., Chen, L.-F., Chang, C.-C., & Chiu, F.-H. (2016). Configurational path to customer satisfaction and stickiness for a restaurant chain using fuzzy set qualitative comparative analysis. Journal of Business Research, 69, 2939–2949.

Huarng, K.-H., & Roig-Tierno, N. (2016). Qualitative comparative analysis, crisp and fuzzy sets in knowledge and innovation. Journal of Business Research, 69, 5181–5186.

Jiang, C., Wang, Z., Wang, R., & Ding, Y. (2018, July). Loan default prediction by combining soft information extracted from descriptive text in online peer-to-peer lending. Annals of Operations Research, 266(1-2), 511-529.

Khan, N., & Haque, F. (2013, 01 31). Fuzzy based decision making for promotional marketing campaigns. International Journal of Fuzzy Logic Systems, 3(1).

Korol, T. (2019, January 01). Fuzzy Logic in Financial Management. Retrieved from IntechOpen: https://www.intechopen.com/books/fuzzy-logic-emerging-technologies-and-applications/fuzzy-logic-in-financial-management

Kraus, S., Soriano, D., & Schüssler, M. (2018). Fuzzy-set qualitative comparative analysis (fsQCA) in entrepreneurship and innovation research – the rise of a method. International Entrepreneurship Management Journal, 14, 15-33.

Lending Club. (2019, January 16). Retrieved from LendingClub: https://www.lendingclub.com

Lin, X., Li, X., & Zheng, Z. (2017). Evaluating borrower's default risk in peer-to-peer lending: evidence from a lending platform in China. APPLIED ECONOMICS, 49(35), 3538–3545.

Lisboa, A., Skarmeas, D., & Saridakis, C. (2016). Entrepreneurial orientation pathways to performance: A fuzzy-set analysis. Journal of Business Research, 69, 1319–1324.

Malhotra, R., & Malhotra, D. K. (2002). Differentiating between good credits and bad credits using neuro-fuzzy systems. European Journal of Operational Research, 136, 190-211.

Mammadli, S. (2016). Fuzzy logic based loan evaluation system. Procedia Computer Science, 102, pp. 495 – 499. Vienna, Austria. doi:10.1016/j.procs.2016.09.433

Mangaraj, B. (2016). Relative effectiveness analysis under fuzziness. 12th International Conference on Application of Fuzzy Systems and Soft Computing. 102, pp. 231 – 238. Vienna, Austria: Procedia Computer Science.

Marqués, A., García, V., & Sánchez, J. (2012). A literature review on the application of evolutionary computing to credit scoring. Journal of the Operational Research Society, 64(9), 1384-1399. doi:10.1057/jors.2012.145

Mendel, J., & Korjani, M. (2012). Charles Ragin's Fuzzy Set Qualitative Comparative Analysis (fsQCA) used for linguistic summarizations. Information Sciences, 202, 1-23.

Menekay, M. (2016). Bank credit authorization using fuzzy expert system. 12th International Conference on Application of Fuzzy Systems and Soft Computing. 102, pp. 659 – 662. Vienna, Austria: ICAFS.

Musayev, A., Madatova, S., & Rustamov, S. (2016). Evaluation of the impact of the tax legislation reforms on the tax potential by fuzzy inference method. 12th International Conference on Application of Fuzzy Systems and Soft Computing. 102, pp. 507 – 514. Vienna, Austria: Procedia Computer Science.

Ntiamoah, E. B., Oteng, E., Opoku, B., & Siaw, A. (2014). Loan Default Rate and its Impact on Profitability in Financial Institutions. Research Journal of Finance and Accounting, 5(14).

Paul, S. (2014). Creditworthiness of a Borrower and the Selection Process in Micro-finance: A Case Study from the Urban Slums of India. Margin: The Journal of Applied Economic Research, 8(1), 59-75.

Pearson, Jr., R., & Greeff, M. (2006). Causes of Default among Housing Micro Loan Clients. Siana Strategic Advisors (Pty) Ltd., Michael Greeff Business Consulting cc. Johannesburg, Republic of South Africa: FinMark Trust, Rural Housing Loan Fund, National Housing Finance Corporation, Development Bank of Southern Africa. Retrieved October 21, 2018, from https://housingfinanceafrica.org/app/uploads/COD_final_report.pdf

Poorkavoos, M., Duan, Y., Edwards, J., & Ramanathan, R. (2016). Identifying the configurational paths to innovation in SMEs: A fuzzy-set qualitative comparative analysis. Journal of Business Research, 69, 5843–5854.

Predicting loan defaults. (2018, October 20). Retrieved from Google Site: https://sites.google.com/site/sramadosloan/

Ragin, C. (2000). Fuzzy-set social science. Chicago: University of Chicago Press.

Ragin, C. (2008). Redesigning social inquiry: Fuzzy sets and beyond.

Ramazanov, D., Jabiyeva, A., & Amirguliyev, V. (2016). Fuzzy rule base model for oil wells efficiency estimation. 12th International Conference on Application of Fuzzy Systems and Soft Computing. 102, pp. 198 – 201. Vienna, Austria: Procedia Computer Science.

Rihoux, B., & Ragin, C. (2009). Configurational Comparative Methods: Qualitative Comparative Analysis (QCA) and Related Techniques. Thousand Oaks, CA: Sage. doi:http://dx.doi.org/10.4135/9781452226569

Romaniuk, S., & Hall, L. (1992). Decision making on creditworthiness, using a fuzzy connectionist model. Fuzzy Sets and Systems, 48, 15-22.

Saunders, A., & Cornett, M. M. (1999). Financial Institutions Management: A Risk Management Approach. Boston, MA, USA: McGraw-Hill/Irwin.

Schneider, C., & Wagemann, C. (2010). Standards of good practice in qualitative comparative analysis (QCA) and fuzzy-sets. Comparative Sociology, 9(3), 397-418.

Shmueli, G., & Koppius, O. (2011, September). Predictive Analytics in Information Systems Research. MIS Quarterly, 35(2), 553-572.

Sjödin, D. R., Parida, V., & Kohtamäki, M. (2016). Capability configurations for advanced service offerings in manufacturing firms: Using fuzzy set qualitative comparative analysis. Journal of Business Research, 69, 5330–5335.

Skarmeas, D., Leonidou, C., & Saridakis, C. (2014). Examining the role of CSR skepticism using fuzzy-set qualitative comparative analysis. Journal of Business Research, 67, 1796–1805.

Student loan default in the United States. (2018, October 21). Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Student_loan_default_in_the_United_States

TMR. (2019, January 21). peer to peer lending market. Retrieved from Transparency Market Research: https://www.transparencymarketresearch.com/peer-to-peer-lending-market.html

Tsabadze, T. (2017). Assessment of credit risk based on fuzzy relations. In K. Ntalianis (Ed.), AIP Conference Proceedings. 1836. Rome, Italy: American Institute of Physics.

Tsorhe, J. S., Aboagye, A., & Kyereboah-Coleman, A. (2019, February 1). Corporate Governance and Bank Risk Management in Ghana. Retrieved from Semantic Scholar: https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=2ahUKEwjk3cag0LPgAhXtxMQBHe6CDSoQFjAAegQIBRAC&url=https%3A%2F%2Fpdfs.semanticscholar.org%2F420e%2F535a5c6b4def3494ae27d4d57ed145091ed3.pdf&usg=AOvVaw0RpUJxO4qqG8ISBZGVZ-lH

Tóth, Z., Henneberg, S., & Naudé, P. (2017). Addressing the 'Qualitative' in fuzzy set Qualitative Comparative Analysis: The Generic Membership Evaluation Template. Industrial Marketing Management, 63, 192–204.

Wang, D. H.-M., Yu, T. H.-K., & Chiang, C.-H. (2016). Exploring the value relevance of corporate reputation: A fuzzy-set qualitative comparative analysis. Journal of Business Research, 69, 1329–1332.

Wang, H., Chen, K., Zhu, W., & Song, Z. (2015). A process model on P2P lending. Financial Innovation, 1(3), 1-8.

Vis, B. (2012). The Comparative Advantages of fsQCA and Regression Analysis for Moderately Large-N Analyses. Sociological Methods & Research, 41(1), 168–198.

Woodside, A., & Zhang, M. (2013, February 05). Cultural diversity and marketing transactions: Are market integration, large community size, and world religions necessary for fairness in ephemeral exchanges? Psychology and Marketing, 30(3), 263-276.

Xia, Y., Liu, C., & Liu, N. (2017). Cost-sensitive boosted tree for loan evaluation in peer-to-peer lending. Electronic Commerce Research and Applications, 24, 30-49.

Zhao, H., Ge, Y., Liu, Q., Wang, G., Chen, E., & Zhang, H. (2017, July). P2P Lending Survey: Platforms, Recent Advances and Prospects. ACM Transactions on Intelligent Systems and Technology, 8(6), Article 72: 1-28. Retrieved from https://www.researchgate.net/publication/318665428_P2P_Lending_Survey_Platforms_Recent_Advances_and_Prospects

APPENDICES

Necessity conditions for Scenario 2 (fsQCA)

		inclN	RoN	covN
1	int_rate	0.325	0.526	0.056
2	INT_RATE	0.675	0.518	0.112
3	dti	0.491	0.491	0.078
4	DTI	0.509	0.553	0.091
5	revol_util	0.194	0.761	0.065
6	REVOL_UTIL	0.806	0.269	0.091
7	emp_length	0.446	0.587	0.087
8	EMP_LENGTH	0.554	0.457	0.083
9	home_ownership	0.822	0.206	0.086
10	HOME_OWNERSHIP	0.178	0.821	0.078
11	int_rate*dti	0.267	0.634	0.059
12	int_rate*DTI	0.254	0.692	0.067
13	INT_RATE*dti	0.421	0.660	0.098
14	INT_RATE*DTI	0.452	0.663	0.105
15	int_rate*revol_util	0.118	0.810	0.050
16	int_rate*REVOL_UTIL	0.294	0.618	0.062
17	INT_RATE*revol_util	0.163	0.852	0.086
18	INT_RATE*REVOL_UTIL	0.599	0.570	0.110
19	int_rate*emp_length	0.208	0.734	0.063
20	int_rate*EMP_LENGTH	0.242	0.669	0.059
21	INT_RATE*emp_length	0.364	0.731	0.106
22	INT_RATE*EMP_LENGTH	0.438	0.670	0.104
23	int_rate*home_ownership	0.319	0.550	0.058
24	int_rate*HOME_OWNERSHIP	0.123	0.847	0.064
25	INT_RATE*home_ownership	0.621	0.547	0.109
26	INT_RATE*HOME_OWNERSHIP	0.172	0.845	0.087

27	dti*revol_util	0.161	0.800	0.065
28	dti*REVOL_UTIL	0.440	0.572	0.083
29	DTI*revol_util	0.143	0.839	0.071
30	DTI*REVOL_UTIL	0.476	0.594	0.094
31	dti*emp_length	0.307	0.703	0.082
32	dti*EMP_LENGTH	0.352	0.642	0.079
33	DTI*emp_length	0.307	0.736	0.092
34	DTI*EMP_LENGTH	0.370	0.670	0.089
35	dti*home_ownership	0.481	0.505	0.079
36	dti*HOME_OWNERSHIP	0.165	0.832	0.078
37	DTI*home_ownership	0.496	0.565	0.091
38	DTI*HOME_OWNERSHIP	0.168	0.834	0.080
39	revol_util*emp_length	0.128	0.854	0.070
40	revol_util*EMP_LENGTH	0.141	0.828	0.066
41	REVOL_UTIL*emp_length	0.393	0.656	0.091
42	REVOL_UTIL*EMP_LENGTH	0.488	0.554	0.088
43	revol_util*home_ownership	0.189	0.771	0.066
44	REVOL_UTIL*home_ownership	0.702	0.367	0.091
45	REVOL_UTIL*HOME_OWNERSHIP	0.172	0.830	0.080
46	emp_length*home_ownership	0.426	0.608	0.087
47	emp_length*HOME_OWNERSHIP	0.126	0.879	0.082
48	EMP_LENGTH*home_ownership	0.502	0.519	0.084
49	EMP_LENGTH*HOME_OWNERSHIP	0.157	0.841	0.078
50	int_rate*dti*revol_util	0.108	0.832	0.052
51	int_rate*dti*REVOL_UTIL	0.248	0.686	0.064
52	int_rate*DTI*revol_util	0.098	0.862	0.057
53	int_rate*DTI*REVOL_UTIL	0.243	0.716	0.069
54	INT_RATE*dti*revol_util	0.142	0.867	0.083

55	INT_RATE*dti*REVOL_UTIL	0.390	0.686	0.098
56	INT_RATE*DTI*revol_util	0.132	0.878	0.085
57	INT_RATE*DTI*REVOL_UTIL	0.429	0.681	0.105
58	int_rate*dti*emp_length	0.183	0.773	0.065
59	int_rate*dti*EMP_LENGTH	0.209	0.725	0.061
60	int_rate*DTI*emp_length	0.174	0.803	0.070
61	int_rate*DTI*EMP_LENGTH	0.205	0.755	0.067
62	INT_RATE*dti*emp_length	0.274	0.782	0.098
63	INT_RATE*dti*EMP_LENGTH	0.317	0.739	0.095
64	INT_RATE*DTI*emp_length	0.283	0.788	0.103
65	INT_RATE*DTI*EMP_LENGTH	0.338	0.739	0.101
66	int_rate*dti*home_ownership	0.266	0.639	0.060
67	int_rate*dti*HOME_OWNERSHIP	0.118	0.854	0.065
68	int_rate*DTI*home_ownership	0.254	0.694	0.067
69	int_rate*DTI*HOME_OWNERSHIP	0.118	0.858	0.066
70	INT_RATE*dti*home_ownership	0.417	0.663	0.098
71	INT_RATE*dti*HOME_OWNERSHIP	0.159	0.855	0.086
72	INT_RATE*DTI*home_ownership	0.445	0.667	0.105
73	INT_RATE*DTI*HOME_OWNERSHIP	0.163	0.854	0.087
74	int_rate*REVOL_UTIL*emp_length	0.193	0.771	0.067
75	int_rate*REVOL_UTIL*EMP_LENGTH	0.226	0.716	0.064
76	INT_RATE*revol_util*emp_length	0.114	0.897	0.087
77	INT_RATE*revol_util*EMP_LENGTH	0.125	0.882	0.083
78	INT_RATE*REVOL_UTIL*emp_length	0.335	0.752	0.105
79	INT_RATE*REVOL_UTIL*EMP_LENGTH	0.405	0.694	0.104
80	int_rate*revol_util*home_ownership	0.117	0.813	0.051
81	int_rate*REVOL_UTIL*home_ownership	0.291	0.626	0.063
82	int_rate*REVOL_UTIL*HOME_OWNERSHIP	0.120	0.855	0.066

83	INT_RATE*revol_util*home_ownership	0.161	0.853	0.086
84	INT_RATE*REVOL_UTIL*home_ownership	0.563	0.589	0.108
85	INT_RATE*REVOL_UTIL*HOME_OWNERSHIP	0.167	0.851	0.087
86	int_rate*emp_length*home_ownership	0.207	0.738	0.064
87	int_rate*EMP_LENGTH*home_ownership	0.240	0.679	0.061
88	int_rate*EMP_LENGTH*HOME_OWNERSHIP	0.111	0.863	0.065
89	INT_RATE*emp_length*home_ownership	0.355	0.735	0.105
90	INT_RATE*emp_length*HOME_OWNERSHIP	0.123	0.893	0.089
91	INT_RATE*EMP_LENGTH*home_ownership	0.419	0.680	0.103
92	INT_RATE*EMP_LENGTH*HOME_OWNERSHIP	0.152	0.861	0.086
93	dti*revol_util*emp_length	0.113	0.870	0.069
94	dti*revol_util*EMP_LENGTH	0.124	0.848	0.065
95	dti*REVOL_UTIL*emp_length	0.281	0.740	0.086
96	dti*REVOL_UTIL*EMP_LENGTH	0.327	0.684	0.083
97	DTI*revol_util*emp_length	0.103	0.890	0.074
98	DTI*revol_util*EMP_LENGTH	0.116	0.869	0.070
99	DTI*REVOL_UTIL*emp_length	0.291	0.756	0.094
100	DTI*REVOL_UTIL*EMP_LENGTH	0.352	0.694	0.091
101	dti*revol_util*home_ownership	0.160	0.803	0.065
102	dti*REVOL_UTIL*home_ownership	0.435	0.578	0.083
103	dti*REVOL_UTIL*HOME_OWNERSHIP	0.160	0.841	0.079
104	DTI*revol_util*home_ownership	0.143	0.840	0.071
105	DTI*REVOL_UTIL*home_ownership	0.466	0.602	0.093
106	DTI*REVOL_UTIL*HOME_OWNERSHIP	0.164	0.840	0.081
107	dti*emp_length*home_ownership	0.304	0.706	0.083
108	dti*emp_length*HOME_OWNERSHIP	0.119	0.885	0.081
109	dti*EMP_LENGTH*home_ownership	0.348	0.648	0.079
110	dti*EMP_LENGTH*HOME_OWNERSHIP	0.147	0.850	0.077
111	DTI*emp_length*home_ownership	0.304	0.739	0.092

112	DTI*emp_length*HOME_OWNERSHIP	0.120	0.887	0.083
113	DTI*EMP_LENGTH*home_ownership	0.364	0.676	0.089
114	DTI*EMP_LENGTH*HOME_OWNERSHIP	0.150	0.851	0.079
115	revol_util*emp_length*home_ownership	0.127	0.856	0.070
116	revol_util*EMP_LENGTH*home_ownership	0.139	0.832	0.066
117	REVOL_UTIL*emp_length*home_ownership	0.380	0.667	0.091
118	REVOL_UTIL*emp_length*HOME_OWNERSHIP	0.122	0.885	0.083
119	REVOL_UTIL*EMP_LENGTH*home_ownership	0.455	0.587	0.088
120	REVOL_UTIL*EMP_LENGTH*HOME_OWNERSHIP	0.153	0.848	0.079
121	int_rate*dti*REVOL_UTIL*emp_length	0.171	0.799	0.068
122	int_rate*dti*REVOL_UTIL*EMP_LENGTH	0.198	0.755	0.065
123	int_rate*DTI*REVOL_UTIL*emp_length	0.167	0.816	0.072
124	int_rate*DTI*REVOL_UTIL*EMP_LENGTH	0.197	0.771	0.069
125	INT_RATE*dti*revol_util*emp_length	0.103	0.905	0.085
126	INT_RATE*dti*revol_util*EMP_LENGTH	0.113	0.891	0.081
127	INT_RATE*dti*REVOL_UTIL*emp_length	0.257	0.796	0.098
128	INT_RATE*dti*REVOL_UTIL*EMP_LENGTH	0.300	0.754	0.096
129	INT_RATE*DTI*revol_util*emp_length	0.096	0.912	0.086
130	INT_RATE*DTI*revol_util*EMP_LENGTH	0.108	0.898	0.083
131	INT_RATE*DTI*REVOL_UTIL*emp_length	0.271	0.797	0.104
132	INT_RATE*DTI*REVOL_UTIL*EMP_LENGTH	0.325	0.750	0.101
133	int_rate*dti*revol_util*home_ownership	0.107	0.833	0.052
134	int_rate*dti*REVOL_UTIL*home_ownership	0.247	0.688	0.064
135	int_rate*dti*REVOL_UTIL*HOME_OWNERSHIP	0.114	0.862	0.066
136	int_rate*DTI*revol_util*home_ownership	0.098	0.863	0.057
137	int_rate*DTI*REVOL_UTIL*home_ownership	0.243	0.717	0.069
138	int_rate*DTI*REVOL_UTIL*HOME_OWNERSHIP	0.115	0.863	0.067
139	INT_RATE*dti*revol_util*home_ownership	0.141	0.867	0.083
140	INT_RATE*dti*REVOL_UTIL*home_ownership	0.388	0.687	0.098

141	INT_RATE*dti*REVOL_UTIL*HOME_OWNERSHIP	0.155	0.860	0.087
142	INT_RATE*DTI*revol_util*home_ownership	0.132	0.878	0.084
143	INT_RATE*DTI*REVOL_UTIL*home_ownership	0.424	0.684	0.105
144	INT_RATE*DTI*REVOL_UTIL*HOME_OWNERSHIP	0.160	0.858	0.088
145	int_rate*dti*emp_length*home_ownership	0.182	0.774	0.065
146	int_rate*dti*EMP_LENGTH*home_ownership	0.209	0.727	0.062
147	int_rate*dti*EMP_LENGTH*HOME_OWNERSHIP	0.106	0.869	0.065
148	int_rate*DTI*emp_length*home_ownership	0.174	0.803	0.070
149	int_rate*DTI*EMP_LENGTH*home_ownership	0.204	0.756	0.067
150	int_rate*DTI*EMP_LENGTH*HOME_OWNERSHIP	0.107	0.871	0.066
151	INT_RATE*dti*emp_length*home_ownership	0.273	0.782	0.098
152	INT_RATE*dti*emp_length*HOME_OWNERSHIP	0.116	0.898	0.089
153	INT_RATE*dti*EMP_LENGTH*home_ownership	0.315	0.740	0.095
154	INT_RATE*dti*EMP_LENGTH*HOME_OWNERSHIP	0.143	0.869	0.085
155	INT_RATE*DTI*emp_length*home_ownership	0.281	0.789	0.103
156	INT_RATE*DTI*emp_length*HOME_OWNERSHIP	0.118	0.898	0.089
157	INT_RATE*DTI*EMP_LENGTH*home_ownership	0.335	0.741	0.101
158	INT_RATE*DTI*EMP_LENGTH*HOME_OWNERSHIP	0.146	0.868	0.087
159	int_rate*REVOL_UTIL*emp_length*home_ownership	0.192	0.772	0.068
160	int_rate*REVOL_UTIL*EMP_LENGTH*home_ownership	0.224	0.720	0.065
161	int_rate*REVOL_UTIL*EMP_LENGTH*HOME_OWNERSHIP	0.108	0.869	0.066
162	INT_RATE*revol_util*emp_length*home_ownership	0.113	0.898	0.086
163	INT_RATE*revol_util*EMP_LENGTH*home_ownership	0.125	0.883	0.083
164	INT_RATE*REVOL_UTIL*emp_length*home_ownership	0.329	0.755	0.105
165	INT_RATE*REVOL_UTIL*emp_length*HOME_OWNERSHIP	0.120	0.896	0.090
166	INT_RATE*REVOL_UTIL*EMP_LENGTH*home_ownership	0.392	0.701	0.103
167	INT_RATE*REVOL_UTIL*EMP_LENGTH*HOME_OWNERSHIP	0.149	0.866	0.087
168	dti*revol_util*emp_length*home_ownership	0.113	0.871	0.069
169	dti*revol_util*EMP_LENGTH*home_ownership	0.123	0.850	0.065

170	dti*REVOL_UTIL*emp_length*home_ownership	0.280	0.742	0.086
171	dti*REVOL_UTIL*emp_length*HOME_OWNERSHIP	0.116	0.890	0.083
172	dti*REVOL_UTIL*EMP_LENGTH*home_ownership	0.325	0.687	0.083
173	dti*REVOL_UTIL*EMP_LENGTH*HOME_OWNERSHIP	0.143	0.857	0.079
174	DTI*revol_util*emp_length*home_ownership	0.103	0.891	0.074
175	DTI*revol_util*EMP_LENGTH*home_ownership	0.115	0.870	0.070
176	DTI*REVOL_UTIL*emp_length*home_ownership	0.289	0.757	0.094
177	DTI*REVOL_UTIL*emp_length*HOME_OWNERSHIP	0.118	0.890	0.084
178	DTI*REVOL_UTIL*EMP_LENGTH*home_ownership	0.348	0.697	0.091
179	DTI*REVOL_UTIL*EMP_LENGTH*HOME_OWNERSHIP	0.147	0.856	0.080
180	int_rate*dti*REVOL_UTIL*emp_length*home_ownership	0.171	0.800	0.068
181	int_rate*dti*REVOL_UTIL*EMP_LENGTH*home_ownership	0.197	0.756	0.065
182	int_rate*dti*REVOL_UTIL*EMP_LENGTH*HOME_OWNERSHIP	0.104	0.875	0.066
183	int_rate*DTI*REVOL_UTIL*emp_length*home_ownership	0.167	0.816	0.072
184	int_rate*DTI*REVOL_UTIL*EMP_LENGTH*home_ownership	0.197	0.772	0.069
185	int_rate*DTI*REVOL_UTIL*EMP_LENGTH*HOME_OWNERSHIP	0.105	0.876	0.067
186	INT_RATE*dti*revol_util*emp_length*home_ownership	0.102	0.905	0.084
187	INT_RATE*dti*revol_util*EMP_LENGTH*home_ownership	0.113	0.891	0.081
188	INT_RATE*dti*REVOL_UTIL*emp_length*home_ownership	0.256	0.797	0.098
189	INT_RATE*dti*REVOL_UTIL*emp_length*HOME_OWNERSHIP	0.113	0.901	0.089
190	INT_RATE*dti*REVOL_UTIL*EMP_LENGTH*home_ownership	0.298	0.755	0.095
191	INT_RATE*dti*REVOL_UTIL*EMP_LENGTH*HOME_OWNERSHIP	0.140	0.873	0.086
192	INT_RATE*DTI*revol_util*emp_length*home_ownership	0.096	0.912	0.086
193	INT_RATE*DTI*revol_util*EMP_LENGTH*home_ownership	0.108	0.898	0.083
194	INT_RATE*DTI*REVOL_UTIL*emp_length*home_ownership	0.269	0.798	0.103
195	INT_RATE*DTI*REVOL_UTIL*emp_length*HOME_OWNERSHIP	0.115	0.900	0.090
196	INT_RATE*DTI*REVOL_UTIL*EMP_LENGTH*home_ownership	0.323	0.751	0.101
197	INT_RATE*DTI*REVOL_UTIL*EMP_LENGTH*HOME_OWNERSHIP	0.143	0.871	0.087