Project 6: Loss Given Default Modeling for Online Microloan

Junjie Lai

Abstract:

Along with the rapidly development of the Internet, online microloans has become a rising industry. Because of the great uncertainty and information asymmetry on the online microloans, credit risk management has become an important part of the online lending platform. In this paper, the real online P2P lending dataset from Prosper will be use to analysis the loss given default (LGD) modeling. Firstly, the characteristics and distribution situation of loss given default (LGD) for online microloan will be presented. Secondly, both parametric and semi-parametric models will be explored to predict loss given default (LGD). For the different models, different dataset are selected. Finally, through testing on the out of time validation sample and comparing the evaluation indexes to find a more accurate and reasonable model. Discovery and application a reasonable quantization model is the key role for improving the overall efficiency of the P2P platform and both creditors and borrowers benefit. At the same time, to support and protect the safety of the entire industry development.

Key words:

Online microloan, Loss given default, Quantile regression model

Chapter one: Introduction

1.1 Background

In 2005, the first P2P platform of the world which named Zopa was born in Britain. The first US public welfare P2P platform released in October of the same year, which aimed to help individuals to make loans to entrepreneurs in developing countries. Afterwards Prosper and Lending Club came out in 2006 and 2007 respectively, they broke out a huge energy and occupied the leading position. Under the supervision of the securitization model, the US P2P development scale and the growth rate kept ahead in the world and leading the industry norms of development. As an emerging field of research, while P2P platform has not yet dominated in the market, along with the rapid development of P2P, the related academic research also increased. The first academic study of P2P lending appeared in 2006 (Hulme & Wright, 2006), when the company Prosper website began to disclose and publish their databases, more and more scholars began to participate in the research, such as: Ravina (2007), Popeand and Syndor (2008), Freedman and Jin (2008), Everett (2008). Among them, the studies of the relationship between lender and borrower were particularly intensive and meticulous. However, from the perspective of the whole economy and social science angle to think about this problem, in order to make a sustainable industrial development, P2P platform must remain innovative and dynamic to avoid being submerged in diverse economic environments. Since the outbreak of the subprime crisis, regardless the small financial institutions or the hundred years old banks, even some government agencies who have declared bankruptcy or closure. According to the Federal Deposit Insurance Corporation (FDIC) data which showed that there were 515 banks failed from 2008 to 2015. The basic reasons of the bank failure were the lack of reserve assets due to lower asset value under the crisis, the assets could not be profitable and the lack of funds prepared which leading to the eventual bankruptcy. From the revision of the Basel Agreement, which proposed the use of internal rating system to determine bank regulatory capital in order to improve the risk sensitivity of regulatory capital. Probability of default (PD) and loss given default (LGD) as the quantitative basis for customer ratings and debt ratings, both also constituted the core variables in the internal rating system. Thus Basel agreement has promoted the construction of the loss given default (LGD) model, more importantly to make a correct evaluation for banks’ loan defaults and losses.

In most actual online microloan environment, Prosper and Lending Club were both two typical companies, those who basically using the same operating mode to match the lender and the borrower through the network. During that process ofP2P platform acted as a financial intermediary, responsible for inspecting borrowers’ creditworthiness, economic conditions, the level, the development prospects and repayment capability. Suppose the transaction is successful, the cooperative bank WebBank would issue the loan. P2P platform earned a service and management fee as a profit. "Efficiency first" in the information society, P2P platform build a win-win model that has been favored more and more by people. Based on the data of 2014, Morgan Stanley predicted that the US online loan market size would reach to 1.9 trillion in 2020. Throughout Prosper and Lending Club database from 2007 to 2014, the total losses and default rates showed a trend of decline, but in the long-term it would not be stable, even default rate would touch above 15%. For each investor, they hope to get ensure the capital safety as well as proper use. The P2P platform needed to establish a more accurate and efficiency loss model to help them to define the borrowing rate and credit rating, in the meanwhile guarantee the safety and maximize the interests of investors.

As a result, for the traditional banking loans services or emerging online microloan, the quantization model for the loss given default or recover rate, not only as a key part for the summary of bank's development history but also played an important part in today’s modern banks and P2P platform in the area of risk management system. In this paper, parametric model and semi-parametric model be used to analysis the loss amount and outstanding, then compared the prediction performance for both types of models. In the past, the analysis of online microloan is to find a kind of conditional expectation, The innovation of this paper is focus on the LGD for online microloans results from the whole conditional distribution. It would be a great leap forward. Using the distribution of different situations, LGD for online microloan has more persuasive power and accuracy. An excellent LGD model has a great help to enhance the level of risk measurement, as well improved their competitiveness and profitability. Meanwhile, LGD model has the universal relevance and reference value, also promoted the development of the world's P2P platform.

1.2 Main research contents

This paper analyzed and drew lessons from various countries to the LGD theoretical research and practical results from modeling states, at the same time, combined with the actual situation of online microloan, research on the construction method of the Prosper's LGD estimation model. Focus on the distribution and accuracy of different models were analyzed, then carried out the valuable information from the modeling contrast analysis. Concretely, this paper be divided into three sections.

(1) Detailed studies of LGD theoretical basis and the existing literatures. Focus on the emphasis about the online microloan from the literatures, stepwise distinguished which characteristics would gave an impact on LGD models for online microloan. Reviewed the relevant literatures, summarized and comment from LGD's basic theory which suggested the influencing factors, distribution, and measurement technology and database construction.

(2)Using Logistic regression, parametric (Linear regression), semi-parametric (Quantile regression) to estimate LGD. Firstly, Logistic regression will be compare with binary quantile model. Secondly, the parametric model was defined as the basic model, four types of transformations be concerned including Fractional logit,Probit,Log-log and Beta distribution. Then quantile regression model be explored as a kind of semi-parametric model to regression the best variable transformation result.

(3)Conducted a comparative analysis among models. Select the alternative variables of the LGD model and optimized the model by stepwise regression gradually. By focusing on the effect of model fitting, reliability and accuracy comparative analysis to define the best prediction performance.

1.3 Main research methods

In this paper, we mainly studied the final effect of different LGD models, this part mainly used the methods of mathematical statistics and econometrics. The main application methods were:

(1)In view of the LGD variable selection, the potential independent variables were the types of characteristics variable. In order to conform to the requirements of the model, the technology of setting dummy variable was adopted. Thus the variable type information was transformed into the statistical model available data information. At the same time, the stepwise regression method was used to select the indexes of various kinds of tests.

(2)In view of the LGD variable test, P value, Standard error, chi-square and Wald test which were used to observe the accuracy of the fitting effect and the reliability.

(3)In view of the LGD statistical model. For the binary recovery rate (RR) dataset, Kolmogorov-Smirnov indicator (KS) is used as criteria for evaluating Logistic and binary quantile regression. For the recovery rate (RR) in the range of 0 to 1, generalized linear regression model be used as a benchmark, through the comparison of Root Mean Square Error (RMSE), observed the accuracy changing if the quantile regression tested.

1.4 Content structure

This paper divided into seven chapters, the first two chapters introduced the research question, literature review. The middle three chapters described the data distribution, descriptive statistics and models and evaluation indicators, then shows models’ analysis specification and results. The last chapter summarized the whole paper and present some suggestions.

Chapter one, introduction. This chapter introduced the research background and significance of this paper, put forward the research questions, the research content, research methods, the potential problems and the innovations.

Chapter two, literature review. This chapter reviewed the theoretical basis for the construction of LGD model. Defined the relevant concepts which involved in this article, clearly to show the meaning of loss given default and the discussion of the relationship between default rate and profit for online micro-loan, also for the influence factors of LGD, distribution characteristics and measuring technology which made a brief review for relevant research documents.

Chapter three, data and descriptive statistics. As Prosper sites provided a huge database, the first variable screening and the process of database description and database construction be expounded.

Chapter four, models and evaluation indicators. Firstly, the general idea LGD model building, made a brief overview for each model. Secondly, four types of parametric models variables transformation: Fractional logit,Probit,Log-log and Beta use to make linear regression. Binary quantile regression model and traditional quantile model will be introduced. Thirdly, two kinds of evaluation indicators Kolmogorov-Smirnov (KS) and Root Mean Square Error (RMSE) will be introduced.

Chapter five, analysis specification and results. Describe the process and results for each models. Contrast the above three models through the effect of fitting, the reliability and accuracy. Using Kolmogorov-Smirnov (KS) and Root Mean Square Error (RMSE) as a criteria to determine the best model.

Chapter six, conclusion and apocalypse. Carrying out the study result through the process of research, also forecast the future direction of research.

1.5 Innovations and potential problems

1.5.1 Innovations

Compared with the same type of research, this paper might be innovative in the following aspects:

(1)The probability default rate and loss given default rate were two dimensional angles of credit risk management, which played an important role in the financial system. For the banking industry probability default research has been a long time history by contrast the late start of the study of loss given default. But they have formed a certain scale and structure. For the online microloan platform that was an emerging industry, risk awareness in the online microloan was ambiguous. This paper not only gave the LGD quantitative models for online microloan but also made comparison, prediction of LGD from a conditional expectation to conditional distribution, finally to find out a more complete system to estimate and forecast the recovery rate.

(2) A small number of scholars have applied the quantile regression model to study the LGD for online microloan. The application of quantile model take a more detail observation of the effects of each variable and greatly improved the accuracy of prediction.

1.5.2 Potential problems

Due to the limited information, time period, research capabilities, this study should be improved in the following aspects:

(1)The database from Prosper that all loans were capped at $35,000 and the period only offered from 36 to 60 months. However, throughout the whole online microloan market, they present a complex diversity. More in-depth analysis be needed if the entire industry be considered about.

(2)The selection of models based on the model which used by the previous LGD analysis in the bank industry. The existence of differences between online microloans and banking industry was obviously. This might reach a certain effect.

Chapter two: Theoretical basis and literature review

2.1 Theoretical basis of loss given default

2.1.1 Concept and selection

The loss given default rate was the percentage of the total amount of risk exposure caused by the loss when the debtor in the event of default, which measured the severity of creditors losses. Loss given default rate also determined the extent of recovery rate of debts, because Loss given default= 1 - Recover rate. Recovery rate referred to the ratio of the amount of debt recovery and risk exposure, as a result of the auction of collateral, enforcement or collection to retrieve funds when the debt default. Therefore, the recovery rate was very low unless there was collateral. That means the loss given default rate would be higher. Once the debtor defaults, the creditors' losses included three aspects: (1) loss of principal, (2) bad debts holding cost, like loan interest income, (3) enforcement fee, which including clearing fees, litigation costs, etc. (Til Schuermann,2004). Loss given default was the core content of banking risk management. P2P platform was a kind of market platform that was generally as same as the traditional banking system, even the challenge for them was not different (Klafft, 2008). Therefore, in this article the analysis method of LGD in banking industry could be transferred to the P2P platform.

2.1.2 Loss given default and economic capital

Generally, capital had two functions. Capitals had continuous financing function, which provided funds for business operations and expansion. More importantly, capital could resist risk that shall bear corresponding business risks. For an enterprise, capital financing was the most main functions, enterprises would lost their competitiveness and viability if there was no adequate capital. For P2P platform, which matching borrowers and lenders to make loans and fulfill in charging interest from the borrowers and paying to the lender. This particular form of management, capital needed to resist risk became the primary function to P2P industry. Economic capital was the basis of risk management and risk measurement in the commercial banks. Economic capital needed to include a variety of unexpected losses, in particular, credit risk, market risk and operation risk. For the P2P industry, market risk and operation risk restricted by external factors, but credit risk could be measured accurately by using the accurate risk factors and the risk loss function. There were three important indicators to assess credit risk, exposure at default (EAD), probability of default (PD) and loss given default (LGD). Through these three factors, we could get the expected loss of assets. EL=EAD\*PD\*LGD. For risk managers, the most important probability was the probability of frequency and severity of loss (Pritchett, 1996). Visible, to analyze the economic capital, loss given default is the central role.

2.2 Literature review

2.2.1 Basic theory for the loss given default

The basis analysis of literature of loss given default and recovery rate could be traced back to 1974. Robert C. Merton published "on the pricing of corporate debt: the risk structure of interest rate”. This study took a loan that created by banks and other creditors to provide a possible credit default company customer service for a fixed period and contract duly repay the principal as the research object, in order to price the loan and determined the interest rate. The main contribution from Merton (1974) who provided a new way of thinking for credit risk, option pricing theory was applied to the valuation of assets and defined the occurs of credit default as a random process and linked to the value of the company. However, the deficiency of this research was not practical observation for the quality of credit assets, thus the application of this model in the empirical application was restricted.

On the basis Merton, Altman (1996) measured the loss given default through the method of statistic of average historical data. Grouhy and Galai (1997) further expressed that Merton (1974) model made it could be directly observed as function of loss given default and recovery rate. So the core part of risk management was simplified to the structural analysis of the probability of default and the loss given default rate. Gupton and Stein (2002) constructed a multi factors analysis model for the Moody's Company. Duffie and Singleton (2003) used the asset valuation method that came from asset pricing theory. Whatever measure loss given default rate by using classification statistics, the primary or advanced regression, or the market price of the asset valuation method the premise was to figure out the probability of loss given default distribution and factors, these properties reflect the characteristics of loss given default.

At present, the research on the loss given default was mainly from two aspects on the theoretical level and the rating agencies. The research results were based on the basic theory, influencing factors, distribution characteristic, measures and database construction. For the P2P platform, loss given default rate research was relatively less, but in the area of developing process, operational structure, lending relationship, profitable business models and default risk control models predecessors have made some contributions.

2.2.2 Impact factors of loss given default

The influence factor of loss given default was one of the research emphases and it was very complicated, which was a key reason for the degree of accuracy for data fitting could not be increased when quantization model be used. Empirical studies showed that LGD was not a deterministic value, which depended on the type of debt, priorities, business cycles, a country's bankruptcy system and legal system, each of which contained many uncertain factors. Therefore, the factors of loss given default were focuses of academic attention. Combined with the results of the studies, the factors that affect the loss given default included several aspects:

(1) Relationship between RR and PD

By observing the relationship between PD and RR, the relationship between RR and LGD could be displayed, because LGD = 1 - RR. Although the calculation of regulatory capital in Basel agreement showed that the loss given default and the probability of default were independent of each other assumed by internal rating based (IRB), a large number of studies have indicated that there was a relationship between PD and RR. Although some scholars have found a positive relationship between RR and PD (such as: Acharya, 2003) or there was no relationship between RR and PD, which was independent of each other (such as: Carey & Gordy, 2001). Most of the research results showed that RR and PD were negatively correlated.

Wilson (1997), Carty (2001), Altman and Brady (2002), Altman and Fanjul (2004) whose research achievements have indicted already: For the purposes of debt, RR and PD are negatively correlated. Gupton, Hamilton and Alexandra (2001) used the Moody database in American between the period from 1989 to 2000, through an empirical study of 181 loans, the results showed that there was a significant negative correlation between RR and PD, and the correlation coefficient was -0.78. Nakshi, Madan and Zhang (2001) improved the simplified model, under the assumption that interest rates were risk-free interest rates by testing the sample data of BBB rated corporate bonds, the results showed that RR and PD there existed a strong negative correlation. The results under the risk-neutral conditions indicated that RR would dropped by one percentage point each when PD increased by four percentage points. Hull and White (2004) investigated collateralized debt obligation (CDO) and credit default swaps (CDS) in derivatives market, the results showed that there was a significant negative correlation between RR and PD. Alman, Brday, Resti and Sironi (2005) analysis showed that the overall RR and PD in the economic system showed a negative correlation.

(2) Debt category and level 《1》

The debts’ priority, seniority and collateral were two important factors that largely determined the level of LGD. The seniority of liquidation was the order of the creditors be rapidly from the residual value of the enterprise when the enterprise bankruptcy liquidation. In developed market economics countries, different financial products have different seniority and formed the relevant legal norms. In addition to the traditional financial collateral and physical collateral, through the continuous financial innovation, the bank has developed a number of effective risk mitigation techniques. By giving it a different LGD data, it should be included in the new capital regulatory framework.

Gupton, Gates and Carty (2000) found the average recovery rate for the high level of protected debts was 70%, but unsecured debt recovery rate was only 52%. Similar conclusions were drawn from the study of the bankruptcy of small enterprises in Sweden. The average recover rate for the high level debt achieved 69% and low level debt was only 2%. By comparison, high level debt recover rate was significantly higher than the low level of debt (Thornburn, 2000). Til Schuermann (2004) counted the recovery rate for different levels of debts through using Moody’s default risk service database. From arrangement for different levels of debt recover rate mean value, results showed that senior secured debt 54.26%, senior unsecured debt 38.71%, Subordinated Debt 34.65%, Junior subordinated debt 14.39%.

(3) Economic cycle

Macroeconomic prosperity or recession had a great impact on loss given default. Frye (2000) investigated Moody’s data and counted the recovery for the different economic cycle, The results showed that the recovery rate of debt during the period of economic prosperity was 1/3 higher than that of the economic recession period.

Carey (1998) observed the Life Thirteen insurance company's private insurance situation, found that the economic recession had a great impact on the distribution of LGD. He used simulation method to statistical analysis of LGD, the results found that at the tail of the LGD distribution (located in the 99.90-99.95% section) the loss rate in a recession at least 50% higher than the expansion period if the level of debts below investment grade. In contrast, the differences among the investment grade debt was relatively small. Upon the whole, the fluctuates of macroeconomic was a systemic risk, which had a higher impact on the lower level of debt instruments.《2》

Schuermann (2004) reached a similar conclusion. After a comparison of the recovery rate for the economic recession period and prosperity period through using Moody's default data, the results showed that in the prosperity period, RR reached 41.39% which was significantly higher than 32.07% in the recession period. Bruche (2009) analysis of the relationship between different macroeconomic variables and the recovery rate in the United States, the pattern of results was remarkably similar. Bellotti and Crook (2012) incorporated macroeconomic variables into an LGD model for credit card loans. The results showed that reducing the observed and predicted LGD difference, macroeconomic variables played an important role. From the investigation of Leow, Mues and Thomas (2013), which focused on the inclusion of macroeconomic variables from two different retail loan LGD models. They found that when macroeconomic variable was included, as a consequence the prediction of residential mortgage LGD could be improved, but for personal loans LGD which was not obviously.

(4) Industry factors

Because debtors existed in different industries, it could has a substantial impact on the level of recovery rate. Some studies suggested that the impact of industry conditions on the LGD was not obviously. For example, Gupton, Gates and Garty (2005) did not find the impact of industry conditions on LGD by researching Moody's credit loans. On the contrary, some researchers showed that the industry had a significant impact on loss given default. Under the same set of the other factors, different industries tend to vary the size of the LGD.

Altman and Kishore (1996) classified RR according to the industry perspective, results showed that in the physical asset intensive industries, the RR level of corporate bonds was relatively high. Specifically statistical results showed that the average recovery rate of public utilities reached 70%, the service industry's RR was 46%, the financial institutions was 36%, while the hotel's RR was only 26%. Grossman et al. (2001) research results also confirmed the influence of industry factors on RR. First they differentiates industries into three categories: asset-intensive industries, service oriented and retail respectively and then statistics average loans and bonds RR for each of categories. The results appeared in the statistics loans' recover rate, the average RR in asset-intensive industries reached 95% and retail industry RR was 89%, while service-oriented industries RR was only 42%; For bonds market also had the same characteristics, the level of recovery rate of bonds was to be less than the loans. For example, asset-intensive industries was 60%, service-oriented industries RR was only 3%.

Bharath, and Srinivasan Acharya (2003) investigated the problems of the industry dilemma, that was the influences of RR appeared when the entire industry downturn. They found that when an industry got into trouble, the mean of LGD for corporate defaults debts was 10% to 20% higher than that when an industry was not in trouble. At the same time, when an industry got into trouble, the expectations for this industry would be deteriorated thus affect the market transaction price for the company's debt, which lead to increase the level of LGD when the default occurs.

(5) Scale factors

Size scale was the decisive factor to construct probability default model. But whether there was a significant impact on the estimation of LGD, there was no conclusive.

Asarnow and Edwards (1995) investigated Citibank’s 831 medium and large companies defaulted loans project between 1970 and 1993, the results showed that there was no significant relationship between LGD and loans’ scale.

Eales and Bosworth (1998) studied through the Australian small business loans and a large amount of consumer loans found the size scale of the loans have an impact on the LGD. The results showed that distribution of RR for the default loans was U shape, In other words, compared with the medium scale loans, RR for small scale and large scale loans was higher.

Garty and Lieberman (2006) researched on the moody’s syndicated loans data, the results showed that there was a negative correlation between the size scale and LGD.

2.2.3 Online lending characteristics

After the 2008 financial crisis, with the economic downturn in traditional financial institutions the on line financial began to develop rapidly long with the popularization and rapid development of the Internet. Internet lending has been favored by people. Slavin (2007) found net lending in the United States and the United Kingdom to become the most important way of financial management after traditional savings and investment to the people. The characteristics and influencing factors of net loan which were paid attention to by the scholars.

In the process of online loans, the levels of interest rate willing to pay was the most important factor that determined whether the borrower could succeed in obtaining a loan, following was the credit rating by the borrowers. Another factor that influenced the successful of lending was the proportion of the loan to the income ( Klafft,2008, Lyer et al.,2009 ).In the borrower's anonymous network environment, the loan risk was higher than the traditional lending (Klafft 2008).

Ravina (2007), Popeand and Syndor (2008) found that determined a successful financing influence factors were the characteristics of financing, such as race, age, gender, weight, beautiful appearance which were obviously affect the successful financing and financing rates. Meanwhile, blacks had relative higher borrowing risks, costs and default rates compared to other people.

Freedman and Jin (2008), Everett (2008) found that if the borrower or lender in real life contact or understanding, the loan default probability would be significantly reduced. Viswanathan and Siva (2009) found that the more extensive the borrower social relationship, the higher loan success rate, the lower default rate and lower cost.

Klafft (2008) by analyzing the lending rules of P2P network, the results showed that P2P network had a lot of similarities with the traditional banking system. Klam (2008) for a more detailed analysis of the credit rating, borrowers with low credit ratings were unable to get loans from traditional banks, and in the P2P lending platform it was also difficult to get loan.

Some views suggested that some small and medium enterprises have to accept higher interest rates loans than expectation, because of the asymmetry information (James and Hadlock 2002). But P2P platform provided a lower cost of opportunity for small and medium enterprises (Schenone 2004). In addition, in reducing the asymmetry information, P2P platform could provide a more transparent and secure credit transactions for both borrowers and lenders. Beyond that, P2P platform also had strong liquidity, the convenient transaction and without the collateral features.

2.2.3 Distribution characteristics and various transformations

To further study the loss given default, it was not only limited to the point estimation and prediction, but also extended to the sample distribution of the loss given default description and distribution model construction. From the existing research literatures, many researchers have found that LGD probability distribution exhibits a bimodal character, LGD distribution either higher or lower. The bimodal features appeared on both sides of the mean, the level of the average was often not the highest probability of occurrence.

Til Schuermann (2004) analyzed the distribution of recovery rate for Moody’s bonds and loans between 1970 and 2003. LGD showed the obvious bimodal distribution, relatively higher peak LGD was about 70% to 80% and the lower LGD was about 20% to 30%. In this paper, the author tried to find the reasons for bimodal distribution. Through the research on the distribution of LGD for different types of bonds and guarantee, result showed that the bimodal distribution only appeared after superposed several types of debt instruments. Schuermann (2003), Renault and Scaillet (2004) concluded in the market a pivotal characteristic of the recovery rate distribution was the high concentration of data at total recovery and loss.

In recent years, a wide range of empirical studies showed the beta distribution has been proved to be the most effective for the fitting of the loans’ recovery rate. Micheal (2006) achieved generalized beta regression of the maximum likelihood estimation method through the SAS program. Bruche (2008) introduced the macroeconomics variables and the credit cycle variables into the generalized beta regression model, analysis of the relationship between different macroeconomic variables and recovery rate for the U.S. Bonds.

Based on the empirical evidence, LGD distributions often present a bimodal distribution and the bounded between 0 and 1. Since the method of ordinary least squares (OLS) could range from negative infinity to positive infinity, the fitting degree for linear regression model would decrease. In order to make the model more robust and accurate, various transformations of LGD have been practiced and applied. Gupton and Stein (2002) used beta distribution function to transform the distribution of LGD into a normal distribution and then to model the transformed target with nine factors. Results indicated that beta transformation linear regression gave better predictions than historical average methods. Dermino and Neto De Carvalho (2006) obtained the conclusion that the average recovery conclusion of 71% was in the same order as that received from the US studies, which introduced by a complementary log-log model to forecast the cumulative recover rate of the corporate loans from a Portuguese bank.Bellotti and Crook (2007), and Bastos (2010) used fractional response regression. Crook and Galina (2015) utilized support vector regression (SVR) for predicting the corporate bonds’ loss given default, which indicated that the SVR techniques to make precise predictions for the recovery rate than the linear regression.

2.2.4 Quantitative analysis methods for LGD

(1) Historical data average method

Basel capital agreement allowed the implementation of primary internal ratings based (IRB) method Banks could use average historical loss as predicted value to estimate LGD, mainly because of this method was simple and easy to accept. With the increase of new data accumulated constantly, the historical average loss rate have to recalculate every year to estimate the future LGD.

Historical data average method was carried out weighted analysis, based on historical data of certain types of debt recovery rate. Then calculated the historical average value of the default loss for a certain class or portfolio. At present, this kind of technology was the most widely used method to calculate the loss given default. However, Til Schuermann (2004) pointed out that the method had serious limitations, due to the unique distribution characteristics of LGD, the mean value was not the maximum value of occurrence probability. So the average forecast of historical data may be misleading. In order to overcome the limitation of the historical data average method, Schuermann (2004) put forward the method of asset value estimation that could be more accurate to measure the loss given default.

(2) Mean value table prediction method

Some banks used the expected average table as a debt instrument to estimate the loss given default. This method was generally based on the historical loss given default data and combined with the rating information from the external rating agencies and the opinions of risk management experts which showed the prediction results. In essence, the mean value table prediction method was an improvement on the historical data average method.

(3) Primary model method

The estimation of loss given default could be realized by the traditional regression model. The recovery rate of default debt that was used as the explained variable and defined various possible factors which affect the debt recovery as explanatory variables then the establishment of linear or nonlinear model. This method improved the accuracy by comparing the previous two methods. However, due to the complexity and variability of the LGD factors, the overall prediction results of this method were still poor.

(4) Advanced model method

Due to the limitations of the primary model method, many banks began to study other methods for more accurate estimating the LGD. Nowadays, the most common used method was regression analysis which after transformation methods be used, and which was established by using the least squares method or maximum likelihood method based on the historical data of the debt loss. Qi and Yang (2009) studied LGD of residential mortgages, which claimed that LGD could be illustrated by the linear regression that involved debt characteristics with loan to value acting the single most important role, and using the data set which came from mortgage insurance companies of America (MICA), during the period from 1990 to 2003.

Bastos (2010) found the non-parametric regression tree model to forecast bank loans RR was superior, which compared with the parametric fractional response regression. The results showed that when the out of sample situation was considered, for the shorter horizons the regression tree showed better results for recovery prediction, for the longer horizons the fractional response regression did better performance than the regression tree.

Qi and Zhao (2011) used 3751 defaulted securities in the US from 1985 to 2008 to compare six types of modeling methods for loss given default. Specifically, OLS, fractional response regression (FRR), Inverse Gaussian regression (IGR) and Inverse Gaussian regression with beta transformation (IFR-BT) be used as parametric group, for the non-parametric group, regression tree and neural network be investigated. Results showed that non-parametric model had stronger model fit and accuracy than the parametric model. Crook and Galina (2015) pointed out that the parametric and non-parametric were two types of predictive model which be used in the empirical literature. The most popular models were linear regression models which shown robustness and effectiveness in LGD prediction and explanation. Quantile regression has not yet been widely applied to the field of LGD. Calem and LaCour-Little (2004) predicted the results of mortgage LGD which shown that quantile regression model expressly estimated the realized prices distribution. Then lead to realistic and robust estimated which suitable for using in the Basel II application. Somers and Whittaker (2006) mentioned that quantile regression might be helpful to solve some problems in the retail credit industry. When the distribution was highly non-normal and understanding the changes in the dispersion around the mean was helpful, resulted that forecasting distribution using quantile regression would made a difference, at the same time, quantile regression would provide more comprehensive descriptions of the data than regressions for the mean response.

(5) Other methods

Implicit data analysis, the basic idea of this method was analyzing credit risk information from implied premium rate for a normal bond or loan and then derived the LGD. This approach required the use of sophisticated pricing models and sufficient database. The method has been applied in bond pricing models and credit derivatives. Expert judgment method, this method was mainly based on the expert experience and provided a primary template for the bank in the case of less data. Meanwhile Expert judgment method had a strong subjective judgment and could not confirm its effectiveness.

2.2.5 Empirical study of LGD

Due to the availability and credibility of the data, the research literatures about the loss given default mostly took the American market as the research object. Carty and Lieberman (1996) in the study of Moody's 58 cases of preferential loan from 1989 to 1996, the average recovery rate was 71%, the median was 77%, and the standard deviation was 32%. Hamilton and Carty (1999) took 159 bankruptcy cases as sample, using the market value method found that the average recovery rate was 56.7%,the median was 56%, and the standard deviation was 29.3%. Gupton, Daniel Gates and Carty (2000) took 121 cases of default loans as sample, the results showed that the average LGD value was 69.5% and 52.1% of the bank loans with a senior security and senior unsecured loans respectively. However, the empirical results could not fully support their own situation, and there was a little deviation from the actual situation, this study was the basis of the Losscalc model. Araten, Jocobs and Varshry (2004) studied the history of loan losses in JP Morgan 1982-1999. The conclusion showed that the average accounting LGD was 27% and the average economic LGD was 39.8%.

In addition, some scholars were concerned about the loss given default rates in other countries. Hurt and Felsovalyi (1998) investigated 1149 bank loans in 27 Latin American countries from 1970 to 1996, analyzed the impact of macroeconomic and loan amount on recovery rate, the average recovery rate was 68.2%. Franks (2004) compared the LGD with France, Germany and the United Kingdom, the minimum LGD was in the United Kingdom and the highest in German. Martin Weber (2005) studied 120 companies LGD in German, the average recovery rate was 72.% and standard deviation was 35.46%.

2.3 Chapter summary

This chapter reviewed the literatures review and theoretical foundation of LGD. Emphatically analyzed the influencing factors, distribution characteristics and mathematical model of LGD, which has a great help in the selection of online microloan variables. At the same time also concerned about the online lending characteristics. These reviews which were summarizes and comments for the basis theoretical support for the construction of LGD model in this paper.

Chapter three: Data and descriptive statistics

3.1 Data overview and variable selection

Our dataset obtained from a P2P lending website Prosper.com from the period between 2006 and 2014. The data illustrated more than 200K microloan performance, which includes customers’ total payments and collection information for default loans. Simultaneously, the dataset including 80 variables, which not only displayed loans themselves, such as loans status, borrower rate and term number etc., but also displayed borrowers’ characteristics and behaviors, such as occupation, employment status and monthly income etc. In this paper, the focus point is on the exploring of LGD, a loan was labeled as default (An account failure to make a payment) and charged off (An account miss payments or charged off after 180 days, the company considers bear the loss itself) will be concerned. After screening the whole dataset, a total number of 19811 data remain with the label of default and charged off. According to empirical LGD distributions are often bi-modal and usually bounded between [0, 1] (Crook and Galina, 2015) and some considerations and comparisons between the abovementioned approaches are needed. Recovery rates based on trading prices are overwhelmingly in the range [0, 1], with rare occurrence slightly greater than 1 (Carty and Lieberman, 1996; Gupton and Stein, 2002; Renault and Scaillet, 2004). Then select the data to meet the requirements by formula calculation. RR is calculated as following formulas:

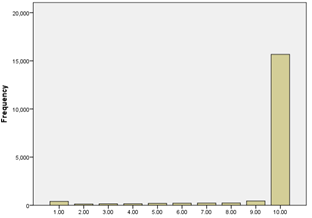
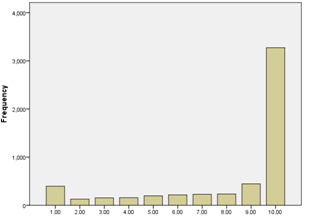
After screened, all the RR which is less than 0 will be excluded and greater than 1 will be adjusted to 0.9999 because of the data error then finally the results show that a total number of 17827data meet the requirements. For the variables selection, first, deleting the variables with a large number of missing data, time period variables and define variables, for example Loan Key, Member Key and Loan origination quarter etc., then through using stepwise test which analyzes the impact of each variables to RR and VIF test which use to avoid the collinear impact. Finally, a total of 35 variables was left, Table 1 shows the detail for those variables.

Table 1: List of variables and their types and explanations

|  |  |  |
| --- | --- | --- |
| Variables name | Type | Explanations |
| 1.EmploymentStatus | String | Employment status of the borrower at the time the listing created. |
| 2.IsBorrowerHomeowner | Boolean | Specifies whether or not the member is a verified homeowner at the time the listing was created. |
| 3.CurrentlyInGroup | Boolean | Specifies whether or not the member is in a group at the time the listing was created. |
| 4.IncomeRange | Interval | The income range of the borrower at the time the listing was created. |
| 5.IncomeVerifiable | Boolean | Specifies whether or not the member’s income is a verifiable at the time the listing was created. |
| 6.ProsperRating | String | The Prosper Rating at the time the listing for the loan was created. |
| 7.ListingCategory | String | The Category of this Listing. |
| 8.EmploymentStatusDuration | String | Duration of the employment status of the borrower at the time the listing was created. |
| 9.CreditScoreRangeLower | Decimal | The lower value representing the range of the borrower’s credit score as provided by a consumer credit rating agency in a recent credit inquiry. |
| 10.CreditScoreRangeUpper | Decimal | The upper value representing the range of the borrower’s credit score as provided by a consumer credit rating agency in a recent credit inquiry. |
| 11.1CurrentCreditLines | Interval | Number of current credit lines at the time the listing was created. |
| 12.TotalCreditLinespast7years | Interval | Number of total credit lines in the last 7 years at the time the listing was created. |
| 13.OpenRevolvingAccounts | Numeric | Number of open revolving accounts. |
| 14.OpenRevolvingMonthlyPayment | Numeric | Number of open revolving monthly payment. |
| 15.InquiriesLast6Months | Interval | Number of inquiries made in the last 6 months. |
| 16.TotalInquiries | String | Total number of inquiries made. |
| 17.CurrentDelinquencies | Interval | Number of current delinquencies at the time the listing was created. |
| 18.1AmountDelinquent | Decimal | The monetary amount delinquent at the time this listing was created. |
| 19.DelinquenciesLast7Years | Interval | Number of delinquencies in the last 7 years at the time the listing was created. |
| 20.PublicRecordsLast10Years | Interval | Number of public records in the last 10 years at the time the listing was created. |
| 21.PublicRecordsLast12Months | Interval | Number of public records in the last 12 months at the time the listing was created. |
| 22.RevolvingCreditBalance | Decimal | The monetary amount of revolving credit balance at the time this listing was created. |
| 23.BankcardUtilizatio | Decimal | The percentage of available revolving credit that is utilized at the time this listing was created. |
| 24.AvailableBankcardCredit | Decimal | Total available credit via bank card. |
| 25.TotalTrades | Integer | Total number of trades. |
| 26.TradesNeverDelinquentpercent | Decimal | Percent of trades never delinquent. |
| 27.TradesOpenedLast6Months | Integer | Total number of trades opened in the last 6 months. |
| 28.DebtToIncomeRatio | Decimal | The debt to income ratio of the borrower at the time the listing for this loan was created. This value is null if the debt to income ratio is not available. This value is capped at 10.01 (so any actual debt to income ratio larger than 1000% will be returned as 1001%). |
| 29.StatedMonthlyIncome | Numeric | Number of stated monthly income. |
| 30.LoanOriginalAmount | Numeric | Number of loan original amount. |
| 31.MonthlyLoanPayment | Decimal | The monthly payment made by the borrower. |
| 32.Recommendations | Numeric | Number of recommendations for borrows. |
| 33.InvestmentFromFriendsCount | Integer | Number of investment from friends. |
| 34.InvestmentFromFriendsAmount | Integer | Amount of investment from friends. |
| 35.Investors | Numeric | Number of investors. |

3.2 Data distribution preliminary analysis

Fig. 1 and Fig. 2 shows the distribution of LGDs with different value range. The vertical axis is the frequency and the horizontal axis shows the intervals. The range of LGD is 0 to 1, LGD is divided into 10 sections, the interval length is the same, number 1 means LGD between 0 and 0.1, and so on, and number 10 means LGD between 0.9 and 1. Fig. 1 shows the distribution of LGD for the whole sample. If the data including large proportion of zeros, in the range of 0.9 to 1 has a very high peak after compared to other ranges. This has also confirms online microloans risk was higher than the traditional lending, which has been mention in the literature part. Given the above, online microloan’s LGD has a different distribution to banks’ LGD. The distribution shows that if an account default, the probability of losing everything is more than 80%.If temporarily ignore the range of 0.9 to 1 because about 70% of account RR is zero and observes the rest of interval, the range of 0 to 0.1 also shows a relatively higher peak. That is why in the quantile models part, both traditional and binary quantile model will be concerned. From Fig.2, LGDs are quite widely spread out in online microloan sample, take up a heavy concentration at both ends in the intervals of [0, 0.1) and(0.9, 1], which has the similar distribution if compare to bank’s LGD. Also compliance the point form Til Schuermann (2004) LGD showed the obvious bimodal distribution.The bi-modal distribution shows that if an account default, the most likely case is losing almost everything, which is about60%, followed by LGDs between 0.8 and 0.9 or losing almost nothing, which is about 8.2% and 7.3% respectively.

Fig.1 Distribution of LGDs between [0,1] Fig.2 Distribution of LGDs between [0,1)

**Intervals Intervals**

Table 2: LGD by closed years, employment status, and income range.

|  |  |  |
| --- | --- | --- |
| Variables | Number of observations | LGD mean (%) |
| Panel A. By closed years |  |  |
| 2006 | 93 | 90.4% |
| 2007 | 1521 | 93.4% |
| 2008 | 2628 | 92.9% |
| 2009 | 3443 | 93.2% |
| 2010 | 1639 | 94.3% |
| 2011 | 975 | 94.6% |
| 2012 | 1739 | 93.0% |
| 2013 | 2693 | 93.5% |
| 2014 | 3096 | 93.3% |
| Panel B. By employment status |  |  |
| Employed | 5865 | 93.1% |
| Full-time | 6852 | 93.5% |
| Not employed | 220 | 92.7% |
| Part-time | 244 | 91.8% |
| Retired | 224 | 94.0% |
| Self-employed | 1180 | 94.1% |
| Other | 3242 | 93.6% |
| Panel C. By income range |  |  |
| $1-24,999 | 1683 | 92.3% |
| $25,000-49,999 | 5772 | 93.5% |
| $50,000-74,999 | 3923 | 93.3% |
| $75,000-99,999 | 1710 | 93.8% |
| $100,000+ | 1488 | 93.5% |
| Other | 3251 | 93.5% |

Table 2 shows the number of three variables (closed years, employment status, and income range) observations and the mean of LGD. Due to the existence of a larger number of online microloans cannot repay when default happen. The mean of LGD keep at a high level with little fluctuation. Panel A shows the observations and LGD mean by closed year, using the quantitative perspective to observation default accounts, there are two rising range (2006 to 2009 and 2011 to 2014). The highest number of default happened in year 2009, the financial crisis in 2008 also influence for online microloans, the highest mean of LGD happened in year 2011. In other words, economics cycle also affect P2P market. From Panel B, LGD for self-employment and retire person have relatively higher level.

3.3 Chapter summary

Through the preliminary analysis of the whole dataset, 35 variables and about 20000 accounts remains after screening. Meanwhile, LGD for online microloan shows a bimodal distribution is confirmed. Because so many accounts cannot repay back anything when default happen, so it is necessary to make a distinction between these accounts, which will be carefully explains in the next chapter.

Chapter four: Models and evaluation indicators

In this section, two models will be introduced one is generalized linear model the other is quantile regression. Previous empirical investigation shows that linear regression models appear to be of comparable predictive accuracy as other more complicated statistical models (Bellotti and Crook, 2012; Qi and Zhao, 2011). In this paper, generalized linear model be used as a benchmark to compare with quantile regression which is useful to forecast distribution when distribution is highly non-normal (Mark and Joe, 2007). From data and descriptive statistics chapter, the characteristic of online microloan is RR appears as a 0 value in the most of time when an account default. So the overall data will be divided into two kinds of forms to consider, one is between 0 and 1, and the other is binary (RR is defined as 0 and (0, 1)). Because the RR needs to be more than 0 in the generalized linear model regression. For the range of RR falls in 0 to 1, four types of data transformation generalized linear model be used to compare with quantile model, root mean square error (RMSE) will be used as an evaluation index, However, for the binary RR dataset, binary quantile model be compared with the logistic regression, Kolmogorov-Smirnov indicator (KS) is used as criteria for evaluating this two models. The following section introduces the models and transformations.

4.1 Parametric model

Generalized linear model is a kind of parametric model, for the particularity of RR bimodal and U-shaped distribution, for the binary RR dataset Logistic regression as a kind of generalized linear regression model is used to analysis. For the data falling in the interval of 0 to 1, linear regression model with four types of transformation are used to analysis. Following four types of data transformation is exploited:

(1) Fractional logit transformation:

In many economic environment, fractional response variables naturally occurring. Papke and Wooldridge (1996) investigated employee participation rates and econometric methods for fractional response variables then in-depth study of generalized linear models from quasi-likelihood and statistics perspectives to robust methods for estimation and inference with fractional response variables. Dermine and de Carvalho (2005) analyzed 374 corporate loans LGD through using fractional response variables. To sum up, the fractional logit model is particularly attractive, since it deals specifically with response variables in the range of 0 to 1, this matches the range of RR, by transforming them into a larger range of values (Bellotti and Crook, 2012).

(2) Probit transformation:

Probit is a common transformation for linearizing sigmoid distributions of proportions (Armitage and Berry, 1994). Andersen and Sidenius (2004) proposed the use of a probit transform of the LGD such that the transformed LGD is normally distributed. The probit transformation guarantees that the LGD stays in the interval [0,1].In the formula, Φ is the cumulative density function of the standard normal distribution and R1,..., Rn are observed RRs taken from the training data. By using Probit transformation, RR can be changed into a normal distribution, which is helpful to our mathematical analysis.

(3) Log-log transformation:

The log transformation applied to adjustment highly skewed distributions to less skew. Log transformation can be more valuable for patterns describe and meet the requirement of inferential statistics. Qi and Zhao (2011) analyzed LGD by using log-log function to build fractional response regression.

(4) Beta transformation:

In this formula, is the cumulative density function of the standard normal distribution then is the inverse standard normal distribution.are positive shape parameters which estimated from training data by using maximum likelihood estimation. Due to Beta distribution is useful for bimodal variable with a U- shaped distribution over the interval 0 to 1 that was widely used to transform RR to model LGD. Gupton and Stein (2005) was successfully modeling RR under Beta distribution in Moody’s KMV software. Bellotti and Crook (2012) researched RR for the credit card default data and presented a pattern for Beta transformation distribution. More importantly, for the two extreme values, it describes the steps of transformations to create the normal distribution.

The linear regression model is given as:

4.2 Semi-parametric model

Quantile regression methods will be studied in this paper as a semi parametric model. The idea of quantile regression was originated in 1760, however, the complexity of this regression method is still a big challenge. Nowadays, the wide use of computer and statistical software makes it easy to fit the quantile regression model. Quantile regression is one of the frontier research direction of econometrics. Koenkel and Pxassett (1978) proposed the idea of quantile regression. It is based on the explanatory variables in the multiple quantile (such as quartile, tenth, percentile) to explain the variables in the conditional distribution of the corresponding quantile equation. Quantile regression does not need to make assumptions about the sequence distribution when regression analysis is carried out, and quantile regression can be used to analyze the difference quantile of the distribution of the dependent variable. Compared with the traditional OLS, quantile regression can describe the statistical distribution of the variables in a more detailed way. Therefore, applying quantile regression method in the area of credit risk measure can avoid complex hypothesis of market data and financial data, and can also examine the structural changes of credit risk in different points.

4.2.1 The basic principle of quantile regression

General linear regression models can be set as follows:

u is the random disturbance

Under the premise of Gauss Markov assumption:

are the coefficient of explanatory variable

Above equations, which is the expression of the mean regression model, is the result of the mathematical expectation on both sides of the equation. Similar to the mean regression model, the median regression model can be set as follows:

The resulting quantile regression model is as follows:

Where y is the dependent variable and is thequantiles,

For the mean regression model, the least squares (OLS) can take to estimate the unknown parameters; for median regression model, the least absolute deviation (or least absolute deviation method LAD) can be used; and for the quantile regression model, by taking a linear programming (LP) method to estimate the minimum weighted absolute deviation to get regression coefficients of explanatory variables. More detail will be expressed as follows:

OLS:

Result:++

LAD:

Result:++

QR:

Results:++

In this paper, traditional quantile regression model be used to study heterogeneity in the effect of non-credit related information on microloan metrics and build a model when RR bigger than 0 and less than 1, for a vector of information x, the following function be used to estimate:

Where Y can be loan related metrics, X is the different of characteristics and individual variables. By estimating for different-quantiles, the heterogeneity in the effects of the variables in online microloans lending can be identified.

4.2.2 Advantages of quantile regression model

(1) Quantile regression model does not need to make any assumptions about the distribution of random error term in the model, which makes the model more robust.

(2) Quantile regression method is not described the connection function result of the relationship between the mean and variance of the dependent variables, so the elastic properties in quantile regression are better.

(3) Quantile regression analysis the different distribution of dependent variables in various quantile , it is possible to provide an accurate analysis of the data, especially when there is abnormal point exist, quantile regression can fit the characteristics of the data itself better.

(4) The parameters estimation from quantile regression have progressive goodness under the large sample theory.

4.2.3 Binary quantile regression model

Due to the weakness of traditional quantile regression model that is cannot solve the response variables are binary. Another semi-parametric model named binary quantile regression is applied in this paper which can be conducive to solve the problem when RR is zero or not, meanswhile fewer assumption on the underlying data needed. Manski (1975) first introduced quantile regression and maximum score estimator as a median estimator. Kordas (2004) outlines the benefits of using Manski's (1975, 1985) binary regression quantiles to provide consistent estimates of the conditional probability of some outcome given an individual's covariates at different points of the distribution.

In the case of a given sample i, binary quantile regression model can be defined as follows:

Where is a continuous variable, which determines the value of the binary variable, is a vector of the different quantilecorresponding to the estimated coefficient.represents a random error term.

Kordas (2006) recommend the use of probabilistic methods to predict binary quantile regression, according to the quantile, and a givenindependent variable to obtain the distribution of , and then get the probability of for 0 or 1, which also means an account recovery rate. Binary quantile regression model be used to estimate the probability of funding for a listing and the probability of default for a matured loan, as traditional quantile regression model, for a vector of information x, the following function be used to estimate:

Benoit and Van (2011) found a posterior Bayesian estimation method for the parameters of the binary quantile regression model. Assuming obey the asymmetric Laplace distribution, in the case of a given sample and quantile, can get the combined posterior density function ofand.

The difference of the prior distribution of the parameter determines the difference of the binary quantile regression model. At present, according to the prior distribution ofBayesian binary quantile regression model can be divided into two categories, one is with adaptive lasso variable selection, the other is without adaptive lasso variable selection. By contrast, the prior distribution of theparameters for without adaptive lasso variable selection is normal distribution. Meanwhile, the posterior distribution can be fitted by MCMC technique, R or SAS software packages have been developed to complete the process.

In this paper, setting the corresponding variable as a binary variable selection

Then the binary mean regression model can be build

In the formula, determined by RR, the value is 0 or 1. define as the explanatory variable andas parameter vector, is a random error term. , n as the number of sample observation.

In order to obtain the parameter estimation β, Benoit and Van (2011), Benoit and Poel (2012) viewpoints and methods will be referenced. Bayesian estimation method is used to estimate the posterior distribution of β and Markov Chain Monte Carlo (MCMC) is used to sample the distribution.

In the binary quantile regression model, random error term distribution is assumed as Laplace distribution:

Then, the potential variables also obey the Laplace distribution

Scale parameter artificially set to 1,. When be estimated at quantile. The quantileof the potential variable can be obtained:

At this time the probability can be expressed as:

Whereis a cumulative distribution function for non-symmetric Laplace variable.

In the case of a given sample of joint posterior density of and can be observed:

Where, is a priori distribution of. is an exponential function. Through using Markov Chain Monte Carlo (MCMC) in the joint posterior density function, function can be solved. Obviously, on one hand, binary quantile regression can reveal the effect of explanatory variables on the distribution of the response variables, and observe the effects of different conditions on the distribution of the variables. On the other hand, the model is more robust than the binary regression model, which can better predict the probability.

4.3 Evaluation indicators

The expected predictive accuracy of the models through using two performance measures: Kolmogorov-Smirnov indicator (KS) and root mean square error (RMSE).

Kolmogorov-Smirnov test is a method of testing a single sample then judge whether that is subject to a particular distribution of a predetermined hypothesis. The test method is based on difference between the cumulative frequency distribution of the sample data and a specific theoretical distribution, if the gap between the two is small, then it can be identified the samples are drawn from a particular distribution. The KS is defined as:

Where is the indicator function, which is equal to 1 if or equal to 0 otherwise.

For a given function, the Kolmogorov-Smirnov statistic can be calculate as following:

Where is the supremum of the set of distances.

Root mean square error (RMSE), which is the square root of the sum squares of the observed value and the true value divide by the number of observations. In the actual measurement, the number of observations n are limited, the true value only can replace by the trust or optimal value. On a set of measurements RMSE reflecting very sensitive whether very large error or especially small error, therefore, RMSE can well reflect the precision of the measurement. The RMSE is defined as:

Where are the actual and predicted recovery rates for an account.

4.4 Chapter summary

Above is the detail introduction for the all models and evaluation indicators for the whole paper. SAS and R software will be used to run these models and generate prediction result, the steps and results will be present and analysis in the next chapter.

Chapter five: Analysis specification and results

This section describes the application process for each models and summarize results from different modeling methods for LGD after comparative analysis. The same data set and independent variables across all models methods. The results for logistic regression model and binary quantile regression model is reported in Table X and Fig. X, results from generalized linear model with four types of transformations and traditional quantile regression model in Table X and Fig. X.

Because there are 80 variables in the overall data, there are remaining 35 variables after preliminary screening. In order to strong the model and more accurately, several variables affecting recovery rate were analyzed by stepwise regression model, through the observation of the chi-square and P-value, then an optimum equation was suggested for the main variables affecting recovery rate, totally have 13 variables. See Appendix For the missing data, this paper makes a fill of them, for the character data, using the mode instead of the missing value, for the numeric data, median be used to instead of missing value. At the same time, clustering the character type variables which have relatively more of classification, such as, clu\_ES(NA&E&R) which is a cluster group and including the employment status with not available, self-employed, retired and employed. More detail shows in See Appendix, reducing the variables is helpful to regression analysis. For the division of training set and validation set, in this paper, based on the origination data of the loan, separate the in sample data set and out of sample data set.

5.1 Logistic regression and binary quantile regression

This part shows the result of Logistic regression model and Binary quantile regression model. Through comparing the KS index for both models, it is can be concluded that binary quantile regression is better than Logistic regression model for the binary RR dataset. The following part is a detailed exposition of two models.

5.1.1 Logistic regression model

Table 3:Result of maximum likelihood estimation for Logistic regression

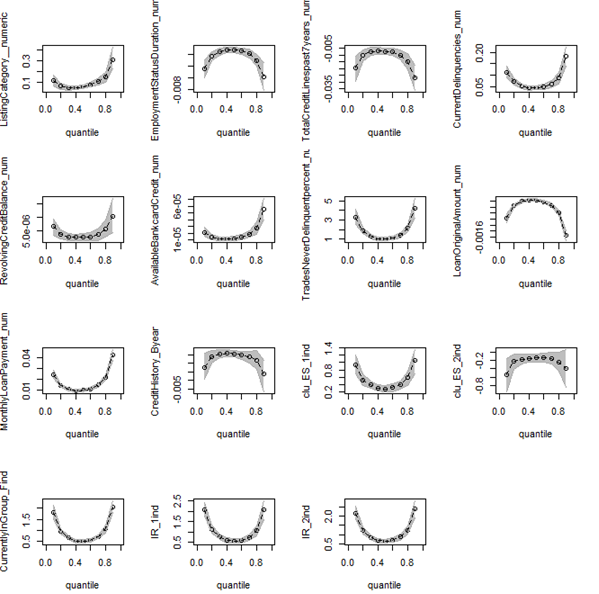
| **Parameters** | **Degree of freedom** | Estimate | **Standard error** | **Wald** | **Pr > Chi-square** |
| --- | --- | --- | --- | --- | --- |
| **Intercept** | 1 | 1.0301 | 0.1340 | 59.1046 | <.0001 |
| **ListingCategory** | 1 | -0.0450 | 0.0117 | 14.8313 | 0.0001 |
| **EmploymentStatusDuration** | 1 | 0.00112 | 0.00031 | 12.9926 | 0.0003 |
| **TotalCreditLinespast7years** | 1 | 0.00563 | 0.00162 | 12.1173 | 0.0005 |
| **CurrentDelinquencies** | 1 | -0.0367 | 0.00605 | 36.7476 | <.0001 |
| **RevolvingCreditBalance** | 1 | -1.97E-6 | 7.651E-7 | 6.6590 | 0.0099 |
| **AvailableBankcardCredit** | 1 | -9.04E-6 | 1.934E-6 | 21.8393 | <.0001 |
| **TradesNeverDelinquentpercent** | 1 | -0.8686 | 0.1273 | 46.5805 | <.0001 |
| **LoanOriginalAmount** | 1 | 0.00029 | 0.00002 | 133.787 | <.0001 |
| **MonthlyLoanPayment** | 1 | -0.00789 | 0.00070 | 126.091 | <.0001 |
| **CreditHistory\_Byear** | 1 | 0.00067 | 0.00027 | 6.2418 | 0.0125 |
| **clu\_ES\_1ind** | 1 | -0.2493 | 0.0558 | 19.9816 | <.0001 |
| **clu\_ES\_2ind** | 1 | 0.1081 | 0.0627 | 2.9697 | 0.0848 |
| **CurrentlyInGroup\_Find** | 1 | -0.4363 | 0.0465 | 87.9757 | <.0001 |
| **IR\_1ind** | 1 | -0.4940 | 0.0676 | 53.4526 | <.0001 |
| **IR\_2ind** | 1 | -0.5546 | 0.0677 | 67.0400 | <.0001 |

Table 3 is the maximum likelihood estimation analysis result in Logistic regression model. There variables have a significant impact on RR.There are seven variables shows a positive correlation with RR, these are Employment status duration (EmploymentStatusDuration\_num), Total credit lines in past seven years (TotalCreditLinespast7years\_num), Loan original amount (LoanOriginalAmount\_num), Credit history by years (CreditHistory\_Byear), Employment status belong to not employment and part time (clu\_ES PT&NE) and Income range between $75,000 and $99,999 and Not displayed accounts (IncomeRange $75,000-99,999 & IncomeRange Not displayed) respectively, the rest of variables have negative correlation with RR.From literature part, a large number of studies have indicated that there was a relationship between PD and RR, current delinquencies (CurrentDelinquencies\_num) is a kind of PD, result shows that current delinquencies take negative impact on RR. In another words, in the area of online microloan industry PD and RR exist negative correlation. From the coefficient of loan original amount (LoanOriginalAmount\_num), when the original amount bigger RR will be higher. However, the higher monthly loan payment (MonthlyLoanPayment\_num) the lower level of RR. There two variables verified the scale factors which has been explain in literature part takes some impacts for RR.But, different variables size take different impacts for RR, in the area of P2P online microloans market, it is hard to determine the impact from loans’ scale or size. In the banking industry, debt category and level is distinguish the debt security level and the degree of protection, which is an important factor for LGD. For the online microloan market, persons’ credit line (TotalCreditLinespast7years\_num), income range (IncomeRange) and credit history (CreditHistory\_Byear) seems to be a kind of personal guarantee and income level as a personal security level. When a person has higher credit lines, higher income range and longer credit history, the RR will be higher.For the two group of employment status (clu\_ES), the result seems unusual, the person belongs to not employment and part-time have positive impact for RR. Conversely, the person belongs toself-employed, employed, retire and not available have negative impact on RR. The only explanation is that perhaps the person belongs to not employment and part-time need a more rigorous review than other person or they have a very good reason to borrow money and higher security. In general, the research on RR characteristics for online microloan and traditional bank exist in certain similarity, but related degree is not very high.

5.1.2 Binary quantile regression

The variables coefficient in binary quantile regression is estimated by R software. Totally there are 15 variables be test which is the same as Logistic regression model. Fig. 3 shows the coefficient for each variables in different quantile points. In the view of overall situation, most of the extreme values are generated in the head or tail, the middle part is relatively stable. It also indirect shows the RR and LGD have a bimodal distribution characteristics. For some variables, although the size of the regression coefficient does not have a marginal impact, but the positive or negative sign can reflect the direction of the influence of the explanatory variables on the RR. Such as, employment status situation which indicates that the different employment status has different effects on RR and there was a significant heterogeneity at the high quantile points.

Fig. 3 Variable coefficient distribution in different quantile



The coefficients have changed a lot in binary quantile regression if compare to logistic regression. In these parameters, only intercept cross the range from negative to positive, There are only five variables coefficients completely in the negative range, which are enployment status duration (EmploymentStatusDuration\_num),total credit lines in the past 7 years (TotalCreditLinespast7years\_num),loan original amount (LoanOriginalAmount\_num), Credit history (CreditHistory\_Byear) and Employment status belong to not employment and part time (clu\_ES PT&NE), which means in the whole of quantile, these variables take negative impact to RR,however, there five variables shows positive relationship with RR in Logistic regression.Why this happen?The rest of the variables have a more or less positive effect on RR.In a word, binary quantile regression model analysis the variables on RR particularity and concretely.

Table 4 Variable coefficient in different quantile

5.1.3 Comparison result

Above is the result of parameters from the process of training sample. In order to compare the accuracy and degree of fitting between two models, validation set variables will be brought into the models and obtain the prediction value of RR, then compare to the real RR. Table 5 and Table

6 shows two models Kolmogorov-Smirnov Two-sample test result respectively.

Table 5: Logistic regression KS test Table 6: Binary quantile regression KS test

|  |  |  |  |
| --- | --- | --- | --- |
| **Kolmogorov-Smirnov Two-sample test** | | | |
| **KS** | 0.021488 | **D** | 0.056984 |
| **KSa** | 1.426942 | **Pr > KSa** | 0.0341 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Kolmogorov-Smirnov Two-sample test** | | | |
| **KS** | 0.018593 | **D** | 0.049308 |
| **KSa** | 1.234732 | **Pr > KSa** | 0.0948 |

Compare the KS value in the two tables, Binary quantile regression model hold lower KS value than Logistic regression model, lower KS means relative small error between the forecast RR and real RR. Meanwhile, lower KS also prove the distribution of the test set and the validation set is more similar. With the above analysis, for the binary RR dataset, Binary quantile regression model can make more accurate predictions then logistic regression model.

5.2Linear regression and quantile regression

In this section, testing RR falls in the range of 0 to 1, totally XXXX number of dataset. For the missing data, the repair method is consistent with the previous. Because the dataset is relatively small, time interval is not enough dispersion, the split of the training set and the validation set through using randomly split. Four types of data transformation be used for the processing of RR, for the different transformation, SAS system through backward test automatically match the best variables then get the regressive equation. Root mean square error (RMSE) as the evaluation index, the best data transformation results be used for quantile regression. Following four transformation result will be showed.

|  |  |  |
| --- | --- | --- |
| Table 8: Log-log transformation result | | |
| Parameter | Estimate | Pr>|t| |
| Intercept | 5.037 | <.0001 |
| IR\_1 | 0.707 | 0.0079 |
| IR\_2 | 0.389 | 0.1522 |
| IncomeVerifiable\_False | -0.652 | 0.0752 |
| ProsperRating | -0.355 | 0.0024 |
| ListingCategory | -0.081 | 0.0183 |
| OpenRevolvingMonthlyPayment | 0.001 | 0.0031 |
| TotalInquiries | -0.017 | 0.0327 |
| CurrentDelinquencies | -0.069 | 0.0199 |
| RevolvingCreditBalance | 0.000 | 0.0602 |
| TradesNeverDelinquentpercent | -2.092 | 0.0001 |
| DebtToIncomeRatio | 0.238 | 0.0719 |
| LoanOriginalAmount | 0.001 | <.0001 |
| MonthlyLoanPayment | -0.036 | <.0001 |
| InvestmentFromFriendsCount | -0.549 | 0.1001 |

|  |  |  |
| --- | --- | --- |
| Table 7: Fractional logit transformation result | | |
| Parameter | Estimate | Pr>|t| |
| Intercept | 3.704 | <.0001 |
| IR\_1 | 1.442 | <.0001 |
| IR\_2 | 1.055 | 0.0009 |
| IncomeVerifiable\_False | -0.687 | 0.1081 |
| ProsperRating | -0.435 | 0.0014 |
| ListingCategory | -0.077 | 0.0517 |
| OpenRevolvingMonthlyPayment | 0.001 | 0.0004 |
| InquiriesLast6Months | -0.065 | 0.0123 |
| CurrentDelinquencies | -0.101 | 0.0036 |
| RevolvingCreditBalance | 0.000 | 0.0183 |
| TradesNeverDelinquentpercent | -2.271 | 0.0004 |
| DebtToIncomeRatio | 0.225 | 0.1441 |
| LoanOriginalAmount | 0.001 | <.0001 |
| MonthlyLoanPayment | -0.039 | <.0001 |
| InvestmentFromFriendsCount | -0.969 | 0.0256 |
| InvestmentFromFriendsAmount | 0.000 | 0.1425 |

|  |  |  |
| --- | --- | --- |
| Table 10: Probit transformation result | | |
| Parameter | Estimate | Pr>|t| |
| Intercept | 0.213 | 0.0761 |
| CurrentlyInGroup\_False | 0.067 | 0.0951 |
| IR\_1 | 0.392 | <.0001 |
| IR\_2 | 0.325 | <.0001 |
| ProsperRating | -0.058 | 0.0089 |
| OpenRevolvingMonthlyPayment | 0.000 | <.0001 |
| InquiriesLast6Months\_num | -0.012 | 0.0051 |
| CurrentDelinquencies | -0.018 | 0.0012 |
| RevolvingCreditBalance | 0.000 | 0.0008 |
| TotalTrades | 0.002 | 0.1171 |
| TradesNeverDelinquentpercent | -0.273 | 0.0088 |
| LoanOriginalAmount | 0.000 | <.0001 |
| MonthlyLoanPayment | -0.004 | <.0001 |
| Recommendations | -0.076 | 0.038 |
| InvestmentFromFriendsAmount | 0.000 | 0.1123 |
| Investors | 0.000 | 0.0544 |

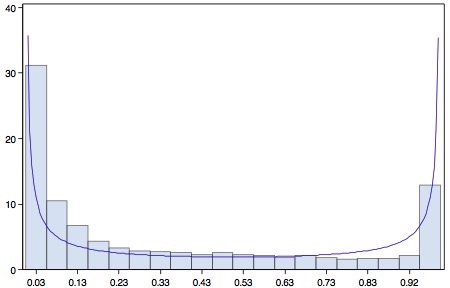
|  |  |  |
| --- | --- | --- |
| Table 9: Beta transformation result | | |
| Parameter | Estimate | Pr>|t| |
| Intercept | 0.409 | 0.0008 |
| IR\_1 | 0.279 | <.0001 |
| IR\_2 | 0.212 | <.0001 |
| ProsperRating | -0.066 | 0.003 |
| ListingCategory | -0.013 | 0.0503 |
| OpenRevolvingMonthlyPayment | 0.000 | 0.0009 |
| InquiriesLast6Months | -0.012 | 0.0055 |
| CurrentDelinquencies | -0.018 | 0.0018 |
| RevolvingCreditBalance | 0.000 | 0.0073 |
| TotalTrades | 0.002 | 0.1561 |
| TradesNeverDelinquentpercent | -0.360 | 0.0007 |
| LoanOriginalAmount | 0.000 | <.0001 |
| MonthlyLoanPayment | -0.006 | <.0001 |
| InvestmentFromFriendsCount | -0.162 | 0.0235 |
| InvestmentFromFriendsAmount | 0.000 | 0.145 |

Observe the above results, although each transformation has a slightly different in variable selection, the core variables near the same if compare with the logistic regression. Such as, income range, revolving credit balance, loan original amount and monthly loan payment etc. In other words, the influence factors of RR or LGD have a certain relationship with the traditional banks. After regression simulation for the validation set, from RMSE results in Table 11, Beta transformation and probit transformation has a better degree of fitting if compare with others. Since probit transformation has some difficulties in the quantile model applications, Beta transformation be choose to apply to quantile model. Another reason to select beta transformation is RR presents a clear bimodal structure when beta be used, Fig. 4 shows RR distribution under Beta transformation and give us a very intuitive bimodal structure feeling.This obeyed the LGD structure has been mentioned in the literature review part and indirectly indicate the reliability of Beta transformation.

Table 11: Four types transformation RMSE result

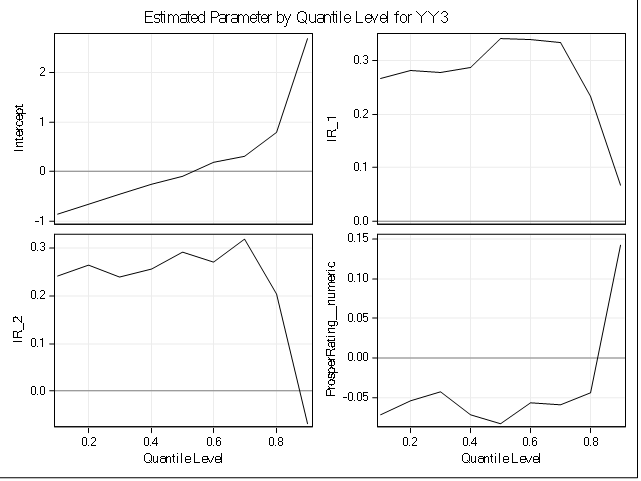
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Transformations | Fractional logit | Log-log | Beta | Probit |
| RMSE | 0.7225 | 0.7394 | 0.6126 | 0.4220 |

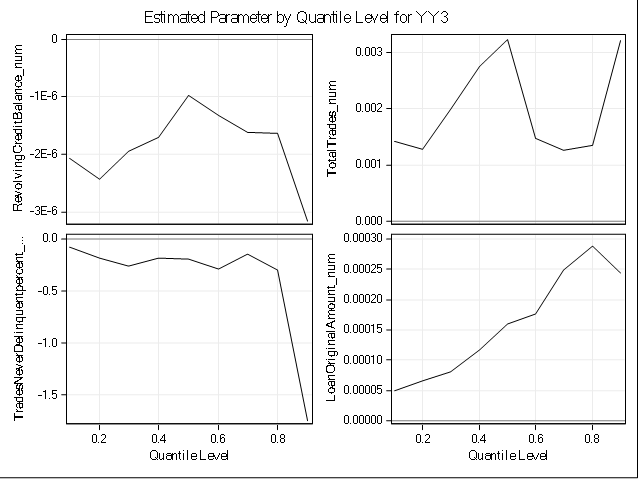
Fig.4: RR distribution under Beta transformation



RR

Using the RR result under Beta transformation, analysis them in the differences quantile through draw into the quantile regression model. The following four pictures shows the confidents of different variables from quantile 0.1 to 0.9. Compare with binary RR quantile result, at this time the coefficients of each variables has a huge wave, it is clear observation of the change in each quantile. In general, most of the extreme values are generated in the tail, because RR presents a special extreme value distribution. More details, in most instances, income range have positive relationship with RR, but when RR fall in the last 30% of the interval, the impact will be significantly decreased and close to 0.For the Prosper rating, listing category, inquiries last 6 months and trades never delinquent percent, in the first 80% intervals, their impact is very smooth, but there is a great change in the last 20% intervals. Monthly loan payment and investment from friends count has negative impact for RR and the impact will be stronger when RR increase. The methods of multiple analytical in various range make quantile regression has stronger analytical skills and each interval has its own characteristics. Table 12 shows the result of RMSE for linear regression under Beta transformation and Probit transformation and quantile regression model, after the observation of the validation set, quantile regression shows the best degree of fitting in the three, which illustrates Somers and Whittaker (2006) viewpoint that quantile regression might be helpful to solve some problems when the distribution was highly non-normal.

Fig.5: Estimated parameter by quantile level for Beta transformation



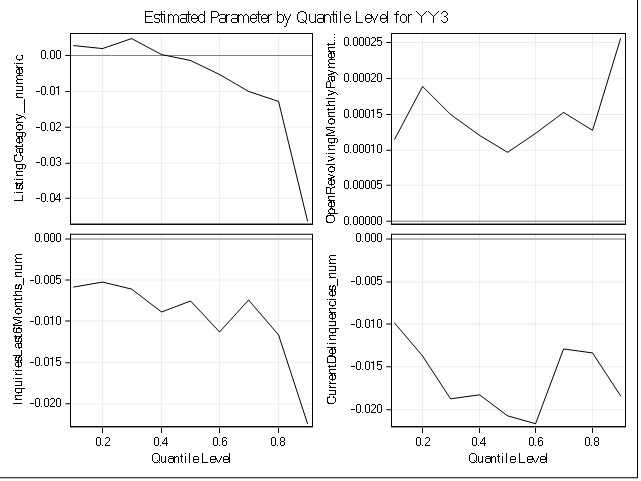
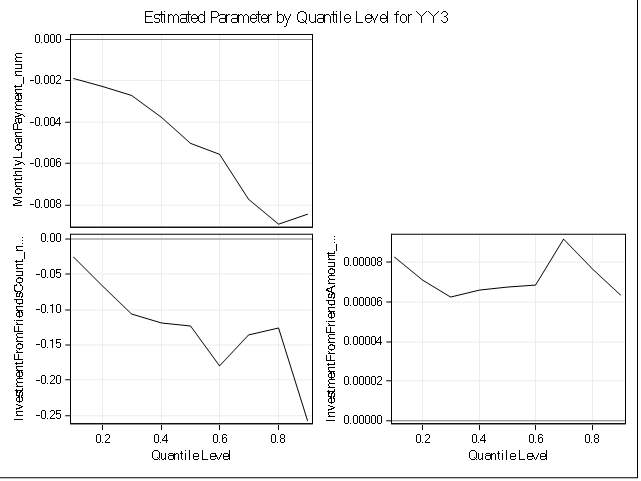


Table 12:Beta/Probit transformation and Quantile regression RMSE result

|  |  |  |  |
| --- | --- | --- | --- |
|  | Beta transformation | Probit transformation | Quantile regression |
| RMSE | 0.6126 | 0.4220 | 0.3437 |

Chapter six: Conclusion and prospected

6.1 Conclusion

Base on the theoretical analysis of the effects of various factors on the heterogeneity of RR, using different models, analysis the degree of fitting of RR in different grouping, especially focus on the quantile regression model. On one hand, through analysis the size and sign of the quantile regression coefficient, reveals the heterogeneity of impact. On the other hand, through the comparison of evaluation indicators KS and RMSE, observation prediction results and degree of fitting. The empirical results shows that quantile regression model has obvious advantages.

Firstly, in general speaking, the LGD distribution on online microloan has a bimodal distribution. There is a certain similarity with the traditional banking, the difference is that the LGD in online microloan have a very high peak in the tail, which means when an account default the totally loss cannot be recover in most cases. The online microloans risk is much higher than the traditional banking. Secondly, for the binary RR dataset, Logistic regression and binary quantile regression be compared. Through observation the evaluation indicators KS, the accuracy of binary quantile regression is better than logistic regression. From the coefficient estimation, the extreme value mostly appear in the header and trailer, the middle part is relatively stable. For the online microloan companies, monitor and tracking these accounts in that ranges can reduce the loss. Thirdly, for the RR dataset fall in the interval 0 to 1, Linear regression with four types of variables transformation and traditional quantile regression be compared. Through observation the evaluation indicators RMSE, quantile regression model shows a lower diversity than linear regression model. From the observation of coefficient estimation, there is a huge mutation in the tails. There is a similarity with the previous results.

Can foreknow, due to the heterogeneity of online microloan behavior, quantile regression model will play a more and more important role in the research of risk, it has confirmed Somers and Whittaker (2006) remarks: when the distribution was highly non-normal and understanding the changes in the dispersion around the mean was helpful, resulted that forecasting distribution using quantile regression would made a difference, at the same time, quantile regression would provide more comprehensive descriptions of the data than regressions for the mean response.

6.2 Prospected

Compared with the classical least square regression, quantile regression is a new and comprehensive statistical method. It can measure the direct relationship between the variables and the regression variables in different quantile, so it has a unique advantage in the application. Quantile regression has only a short history about twenty to thirty years, so a lot of theoretical problems and statistical methods are needed to be further studied. Such as quantile regression time series, number of goodness of fit test and Bayesian quantile regression, they not only well solves the existing quantile regression of some of the problems, but also got the greater attention to this model.

The rapid development of the Internet platform create a good opportunity to online microloan. In order to take more benefits and efficiently on both side of company and customer. On the basis of reasonable operation, they must have a complete set of internal risk analysis system and risk management experience. In order to realize this goal, the combination of the model and the actual situation is particularly important.

In short, in the area of quantile regression theory and P2P online microloans platform, we still need to continue to learn, continue to explore and explain the new problems, to solve the new difficulties, to make us continue to move forward.

Acknowledgments:

At the point of finishing this paper, Please allow me to express my sincere thanks for the people who help me in writing of this paper.

Firstly, I would like to extend my sincere gratitude to my supervisor Dr. Joosung Lee and Dr. Sirong Luo, for their instructive advice and useful suggestions in the process of my writing and thinking.

Secondly, I would like to express my gratitude to Dr. Fernando Moreira, who let me enter the banking and risk field with fresh new eyes and perspective and help me grow up in my master year.

I am also deeply indebted Miss. Xinyu Hu and Miss. Jianron Wang for their help in grammar modification and modeling.

Last but not the least, thanks my parents for their support and encouragement in my study abroad career.

References:

以上引用全要用surname。