HEC MONTRÉAL

Determinants of Repayment Risk in Automobile Loan Market – An Empirical Analysis

By
Mengna WANG

Master of Science in Administration
M.Sc (Finance)

Supervised project presented to the HEC Montréal
in Partial Fulfillment
of the Requirement for the Degree

December 2019

Abstract

As the online collateral lending business has experienced rapid growth in China, credit risk control methods need to be scalable in order to meet growing business demands. This study is based on a real-life business case which features an automobile loan business company wanting to get more insight into the determinants of repayment risk in the automobile loan market so that it can fine-tune the business and reduce management costs.

This work is an attempt to look at the determinants of repayment risk in automobile loans. We examine the historical default risk and prepayment risk in a 4557 auto loans portfolio of a commercial loan company; specifically, we adopt a competing risks framework to estimate Probit and Logit models and then recognize the most important explanatory variables by leveraging some data science approaches.

The results of this study show that the variables traditionally used to predict default and prepayment continue to perform as expected, which includes gender, marital status, loan duration, *LTV* (loan amount/car value) and local economic conditions. We also demonstrate that significant variables that affect default probability and prepayment probability are not the same. However, these two risk probabilities both significatively depend on different time states. In addition, self-selection evidence is also observed: the choice of loan products with different repayment methods and maturities reveals information about a client's propensity to default or prepay. Clients owning automobiles with medium to low-mid brand image (domestic or Korean) have a higher default probability and a lower probability for prepayment. This may refer to a wealth or income effect on default and prepayment probabilities.

Résumé

Étant donné que l'activité de prêt avec collatéral en ligne a connu une croissance rapide en Chine, les méthodes de contrôle du risque de crédit doivent être évolutives afin de répondre à la demande croissante des affaires dans ce domaine. Cette étude est basée sur une analyse d'un cas de la vie-réelle qui présente une entreprise commerciale de prêts automobiles, souhaitant mieux comprendre les déterminants du risque de remboursement sur le marché des prêts automobiles afin de pouvoir affiner les stratégies de l'entreprise et de réduire les coûts de gestion.

Ce travail est une tentative d'examiner les déterminants du risque de remboursement des prêts automobiles. Nous examinons les résultats historiques du remboursement concernant le risque de défaut et le risque de prépaiement dans un portefeuille de 4557 prêts auto d'une société de prêt commercial. En particulier, nous adoptons un cadre de risques concurrents afin d'estimer des modèles Probit et Logit, puis en utilisant certaines approches de la science des données, nous arrivons à reconnaître les variables explicatives les plus importantes.

Les résultats de cette étude montrent que les variables traditionnellement utilisées pour prédire le défaut et le prépaiement anticipé continuent de fonctionner comme prévu, en incluant le sexe, l'état matrimonial, la durée du prêt, le LTV (montant du prêt / valeur de la voiture) et les conditions économiques locales. Nous démontrons également que les variables significatives qui affectent la probabilité de défaut et la probabilité de remboursement anticipé ne sont pas les mêmes. Cependant, ces deux probabilités dépendent toutes deux de manière significative de différents états temporels. En outre, des preuves d'auto-sélection sont également observées: le choix des produits de prêt avec des méthodes de remboursement et des échéances différentes révèle des informations sur la propension d'un client à faire défaut ou à payer d'avance. Les clients possédant des automobiles avec les marques moyenne gamme et bas de gamme (domestiques ou coréennes) ont une probabilité de défaut plus élevée et une probabilité de paiement anticipé plus faible. Cela peut référencer un effet de richesse ou de revenu sur la probabilité de défaut et de prépaiement.

Table of Contents

Table list

Figure list

I. Introduction

Automobile loan assets, mortgage assets as well as financing lease assets are the three big-value property assets in the financial market. Automobile loans have some specific financial characteristics (ex. small-value, short-term and usually collateral-supported); for this reason, it is a business model that can be easily standardized. Furthermore, the business risk of auto loans is lower than other microfinance credit loans or large corporate borrowings, and, in light of the above elements, the industry is enjoying rapid development on a large-scale.

However, as demand for automobile collateral loans is increasing and the domestic credit system is not structured enough to meet this need, the automobile loan industry is facing a great challenge: loan defaults are significant, fraud is frequent, and bad debt rates are high. As a result, practitioners and researchers have been looking for solutions to these problems that seem peculiar to the industry.

At the start of this study, the first question we have to ask concerns the type of repayment risk lenders in the automobile market face. The first and most obvious risk is default risk, that is, people who buy an automobile with a loan or borrow an automobile collateral loan but cannot pay it back. The second major risk for lenders is risk of prepayment, that is, borrowers paying off their loans earlier than the contractual loan's maturity, thereby reducing lenders' interest income.

Over the past two decades, consumer lending has become more complex as lenders shift from traditional interview-based underwriting to data-driven models in order to evaluate default probability and price credit risk. During this transitional period, in the daily lending business of ASFC (Asia Synergy Financial Capital), the risk of default and prepayment are always the main concern of managers. Although they have been looking closely at the problem and have brought many improvements to the risk management process, repayment risk is still not well addressed. For example, risk control before releasing the loan is based essentially on the judgement of the employee who interviewed the client, whereas the company's resource is more focused on risk control after the loan's releasing (ex. collateral assets monitoring.) A lack of risk control from ex-ante is increasing evitable cost and obstructing business development.

In this study, we examine the determinants of repayment risk in automobile loans with the objective of predicting accurately repayment events. Following Agarwal et al. (2008), we adopt a competing risks framework to analyze auto loan default and prepayment risk using a data set of 4557 auto loans signed in ASFC during 2014-10-11 to 2019-08-01. Two baseline Probit and Logit models have been built, and the most important explanatory variables have been selected by leveraging some data science approaches. We hope to apply the findings in this study to complete credit scoring systems in order to improve loan allocation, reduce default risk, and increase the profitability of large auto finance companies. Although the

study was deliberately designed to target only one company, its experiences are certainly not unique. We suspect that some of our findings may be illustrative and drive similar shifts in other companies throughout the industry.

Our main findings can be summarized as follows:

1) Logit and Probit models generate similar significant variables and approximate coefficients. The Logit model has a relatively better predictive power.

2) Significant variables that affect default probability and prepayment probability are not the same. The most important variables selected in the default probability have not necessarily been selected in the prepayment probability equation and vice versa. This implies that limiting the estimation of default probability to calculate repayment risk cost is not enough.

3) Among borrowers' characteristics variables, females have less intention to prepay. Referencing with divorced clients, there is less default probability and higher prepayment intention for married or single clients. Native clients have higher default probability but it is not a significant factor in prepayment model.

4) As for a loan's characteristics variables, longer loan duration increases default probability and decreases prepayment probability. *LTV* (loan amount / car value) rates positively correlate with default probability while negatively correlate with prepayment probability. Borrowers reveal their potential risk exposure through their choice of loan products with different repayment methods and maturities.

5) Among car characteristics variables, clients owning automobiles with a medium to low-mid brand image (domestic or Korean) have a higher default probability and a lower probability for prepayment. If we regard the origin brand image and value of an automobile as a proxy of a client's wealth, this may also refer to a wealth or income effect on default probability. Furthermore, clients who have an auto loan supported by credit [1]have a lower default probability and a higher prepayment probability.

6) Local economic conditions (ex. GDP) are significantly negatively correlated with default probability while being positively correlated with prepayment probability.

7) Default and prepayment probability significatively depend on different time states and interest rates that are positively correlated with default probability.

The rest of this study is broken down as follows: Section 2 introduces the company and business line featured in this project; Section 3 presents a review of literature on the credit scoring of individual risk and the implications of risk management on automobile loan business; Section 4 describes the database used in

[1] The owner of an automobile does not have to collateralize the vehicle to the lender, and the lender controls the risk of auto loans based on the owner's personal information, the owner's credit data and vehicle positioning monitored by GPS.

the study; Section 5 reviews the methodology used for the estimation; Section 6 presents and discusses the univariate and multivariate results obtained; and finally, Section 7 concludes the study.

II. Company profile

Peak Positioning Technology Inc. (Peak) is the parent company of a group of innovative financial technology (Fintech) subsidiaries operating in the Chinese commercial loan industry. As a Canadian-listed company (CSE: PKK), Peak provides a bridge for North American investors, allowing them to participate in the continuous digitalization of Chinese financial services industry and rapidly expanding Fintech sector.

Through its subsidiaries, Peak uses technology, data analyses and artificial intelligence to create an ecosystem of lenders, borrowers, and other participants in China's commercial lending sector. In this ecosystem, the lending business is considered to be is the safest, the most efficient, and the most transparent.

1. Background

The fast development of technology, internet, business intelligence, artificial intelligence, and related technologies have profoundly changed local financial services industries in recent years, and China is not an exception with its nearly thirty years of high economic growth.

China's mobile payment industry has made incredible progress, but the commercial loan industry has not yet fully benefited from these technologies. The financial industry is still operating under the traditional offline brokerage model. Due to the severe shortage of financial services, many SMEs (small & medium-sized enterprises) and individual borrowers cannot find financing channels to conduct business or meet personal financial needs.

The People's Bank of China recently set a minimum limit for small & medium loans to domestic banks. This minimum puts considerable pressure on banks and traditional lenders to be more effective in their lending practices, especially for SMEs; this in turn has created greater opportunities for the proprietary technology introduced by Peak. This technology automates the process by which lenders find and review borrowers. Peak's services enable banks and lenders to increase significantly the amount of loans to meet the required minimum lending targets for SMEs and individual borrowers, which often helps improve the efficiency of China's commercial lending industry. We outline Peak's service delivery and unique business model that is beginning to transform China's business loan industry.

Here are some statistics concerning Chinese commercial lending: the China Association for Small & Medium Commercial Enterprises (CASMCE) reports that by the end of December 2018, there were 43 million SMEs in China. According to CASMCE, a total of $ CAD 25 trillion in loans were made in 2018 and 24% of commercial loans were made to SMEs, micro-enterprises, and individual borrowers. As for the

number of lenders, there were 8351 non-bank lenders in the market as well as 4034 chartered banks. However, it took an average of 1-2 weeks to process and decide on a loan application.

2. Corporate structure and business lines

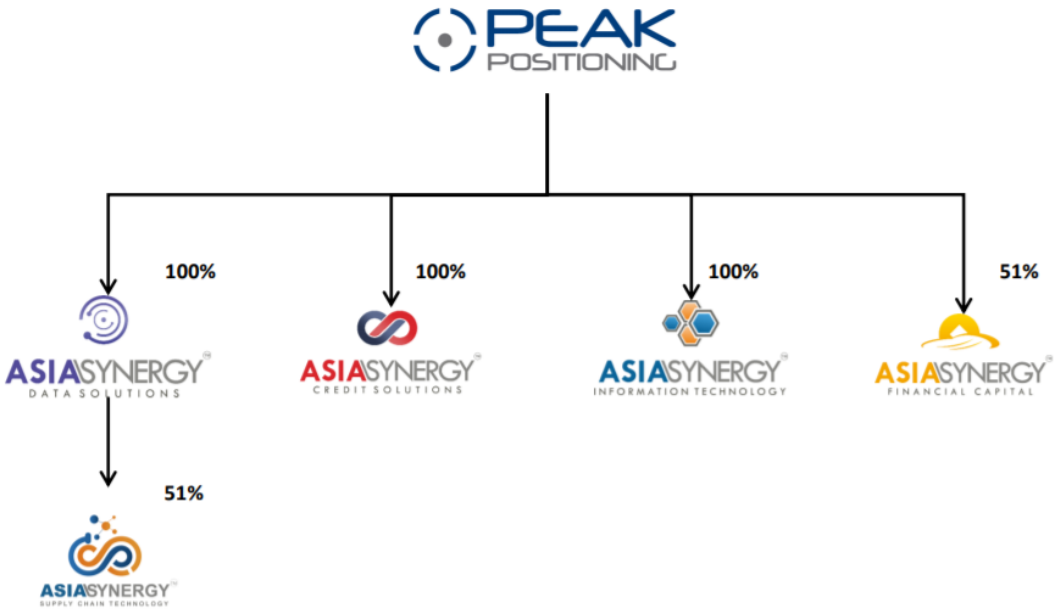
As Peak is a Canadian company that operates in China, its main activity is to provide a bridge for Canadian investors to invest in China's Fintech industry. As of Q3 2019, the company has 713 M shares outstanding and the trading price is between $ 0.065 and $ 0.02. Market Cap is roughly $ 35.6 M at $ 0.05 trading price.

It delivers value to its shareholders by providing services to the Chinese commercial lending industry based on the Cubeler Lending Hub, which is an analytics and artificial intelligence software platform located at the centre of the commercial loan ecosystem, bringing together SMEs, lenders, brokers, data providers, and automated risk management functions to improve the efficiency of commercial loans management. The value proposition for industry participants is:

- **Lenders**: Register on the platform and simply enter their lending criteria. The platform then matches SMEs and individual borrowers to their specified criteria and immediately makes them eligible.
- **SMEs and individual borrowers**: Register on the platform, give access to their business or personal financial data, and get pre-qualified credit from various lenders.
- **Brokers**: Bring their leads to the platform. The platform then qualifies them, matches them with a variety of lenders, and pays an introduction fee for each loan.
- **Data Provider**: Provides data on the platform to generate credit reports, similar to Equifax in Canada.

Each of Peak's five operating subsidiaries plays a specific role and provides different services to industry participants. ASDS (Asia Synergy Data Solution) is the software development and maintenance body for Cubeler Lending Hub. ASFC (Asia Synergy Financial Capital) runs a lending business to SMEs and individual borrowers, with most of these loans being backed by automobile or real estate. ASCS (Asia Synergy Credit Solution) acts as the brokerage business in the group to facilitate big lender operations. ASSC (Asia Synergy Supply Chain Technology) provides logistics, warehousing and other supply-chain related services to manufacturers and their clients and suppliers, including loans and PO (purchase order) financing through a network of financial institution partners. AST (Asia Synergy Information Technology) focuses on supply chain finance and acts as a complementary to ASSC's lending business.

Figure 1. Peak Positioning Inc. - Corporate structure chart



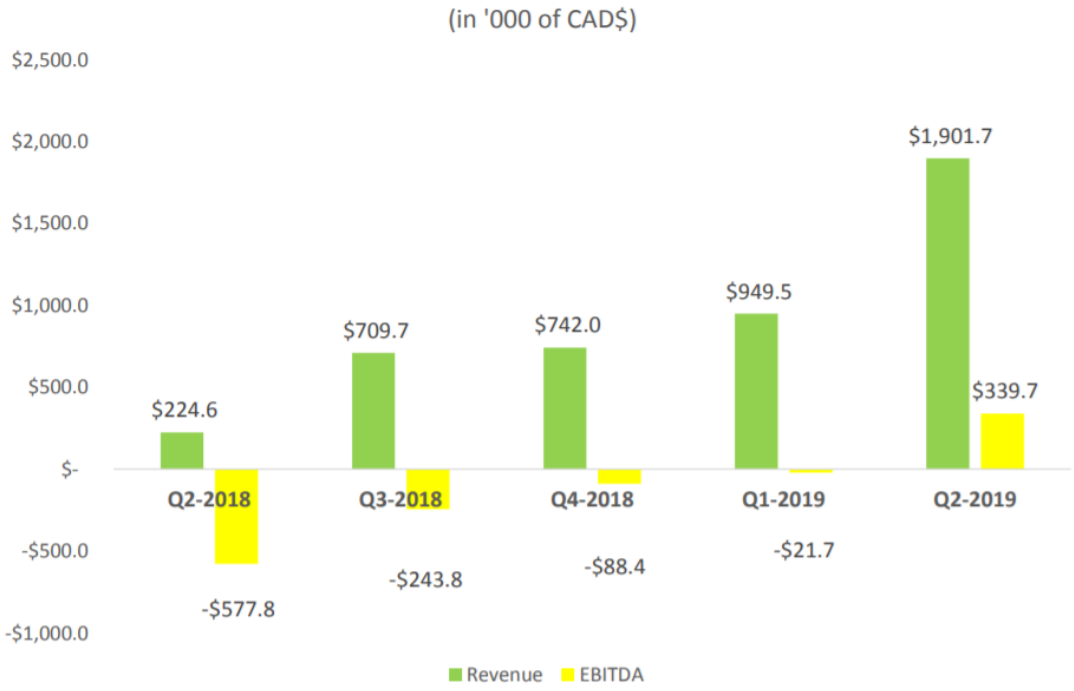
Key Milestones:

- June 2016: Peak established its first subsidiaries in China.
- Mar 2017: Peak established ASDS and obtained exclusive Chinese right to Cubeler platform.
- Aug 2017: ASDS deployed Cubeler Fintech platform in China and recorded first revenues.
- May 2018: Peak established ASFC financial services subsidiary and ASFC made 428 loans worth a combined $ 7.4 M within the first week of operations.
- June 2018: ASFC surpassed 1K loan milestone (existing assets transferred from related party).
- Dec. 2018: Peak established ASCS credit outsourcing services subsidiary.
- Jan. 2019: ASCS signed an agreement with Wuxi Rural Bank to service up to $ 1B in bank loans.
- Jun 2019: Peak established ASSC supply-chain services subsidiary.
- Jun 2019: Peak reported positive EBITDA.

3. Performance review

With a well-established technology and operational infrastructure, Peak has shown five consecutive quarters of steady growth and posted positive EBITDA for the first time in Q2 2019, while continuing to invest in R&D and various business development initiatives. Peak expects that growth trend to continue in 2020.

Figure 2. Peak Positioning Inc. – Performance bar chart



4. Automobile loans business line presentation

Based in Wuxi, which is the capital of Jiangsu province in China, Asia Synergy Financial Capital has operations throughout Jiangsu province and Shanghai. ASFC is a fully licensed financial services institution and provides commercial loans to small and micro-enterprise owners and uses the vehicles of small and micro-enterprise owners as collateral to give out loans. As part of the loan application process, each vehicle is thoroughly inspected, and depending on the use time, condition, and vehicle depreciation, loans issued by ASFC generally do not exceed the fair market value of collateral vehicles. Once the loan application is approved, a patented tamper-resistant smart GPS system is installed on the automobile before the funds are disbursed. ASFC's smart GPS not only allows vehicles to be tracked and located, but also makes them

inoperable remotely. This technology makes ASFC one of the organizations with the lowest loan default rate in the industry. Together with the software to monitor the assets status, ASFC has one of the most sophisticated asset monitoring systems in the vehicle-backed lending industry: ASFC representatives are constantly monitoring the status of the vehicles used to secure its loans by leveraging this technology to ensure it retains visibility. That said, so far, the risk management department of ASFC has focused more on risk control after the loan's releasing; furthermore, risk exposure evaluation and default prediction of the loan applicant at ex-ante is relatively lacking, which increases risk management cost.

5. **Mission description**

My role in the company was to assist the risk management department and work with developers and operations to identify the most important risk factors that impact the lending business. In addition to the above, I built models to examine the data by leveraging recent data science theories.

III. Literature review

Researches on credit scoring and on the elements or factors that would impact the default or prepayment of automobile loan have been closely monitored and discussed over time. Different papers have examined this topic from different angles.

In researching credit scoring for personal loan, Dionne et al. (1996) show that limiting credit scoring to evaluate the default probability without considering different costs and benefits is insufficient to maximize bank profits. In this paper, they estimate jointly the default probability and two conditional truncated distributions of non-payments of good loans and bad loans respectively. They show that the two conditional distributions do not follow the same distribution: the conditional non-payments distribution of defaulters is significative to be a member of the Poisson family distribution, while for defaulters, many no-observed factors still do matter. In estimated models, they use three groups of explanatory variables: personal variables (ex. date of birth, marital status, etc.); socio-economic variables (ex. net monthly income, geographical location, etc.); and financial variables (ex. monthly instalment, availability of credit card, etc.) Finally, they show that the significant variables affecting the three distributions mentioned above are not the same: for example, being in the 25-39 year age group affects the distribution of non-payments; however, it does not affect default probability.

Dionne & Vanasse (1993) propose an empirical analysis of the presence of adverse selection in an automobile insurance market. They show that individuals who choose larger deductibles have an average frequency of accidents lower than those who choose smaller deductibles. However, the authors also mention that this result does not imply adverse selection in the portfolio, because the expected number of accidents is obtained from observable variables. They conclude that there are no remaining adverse choices in the portfolios studied, as adverse selection does not require other optional mechanisms (the deductible choice for example). Insurers can control adverse choices by using appropriate risk classification procedures. In order to obtain the personal accident probability, the authors estimated a regression with parameters such as sex, marital status, age, group of vehicles which are related to personal risk in a latent model and used in predicting this regression to construct the expected number of personal accidents.

D'astous, Dionne and Bergerès (2015) focus on individuals who already have both a revolving line of credit and a term loan. They study the link between the two financial instruments by analyzing both personal credit utilization and personal term loan default probability. To consider the endogeneity of the dependent variables, they model the two dependent variables in a simultaneous equation in which credit line utilization is estimated by instrumental specifications (linear OLS and non-linear regressions (fractional Logit and zero-one inflated beta regressions)) and default probability in the loan is modeled by an instrumented Probit

regression. They find strong evidence of the dependence between the two financial instruments. In more detail, individuals in the default state increase their credit line by 9 percentage points, and according to specifications, an increase of the drawdown rate by 10% of the credit line will reduce the probability of default by 0.09 percentage points to 0.41 percentage points (with a basic default rate of 1.08%). This highlights the fact that borrowers can use the liquidity of credit lines to repay term loans during periods of financial distress and suggests that banks should carefully consider the interactions between a borrowers' credit instrument and manage simultaneously both financial instruments.

In the working paper of Dionne & Liu (2017), the authors provide empirical evidence on the effectiveness of insurance pricing incentives in improving road safety by comparing the frequency of claims after implementation of a regulatory reform in China's pilot cities with the experience of other cities that have not been affected by the reform. They show that the treatment effect of the insurance pricing reform is heterogenous with respect to an insured driver's wealth. Given the fact that the data on income and wealth is unavailable, they consider the value of the vehicle (the purchase price of a new vehicle) as a proxy for the wealth of the insured driver and suggest that less-wealthy insured drivers respond more to the insurance stimulus reform than wealthy insured drivers do. They also find that when considering the origin of the vehicle (domestic and imported) as a proxy for owner's wealth, the impact of insurance incentives on the origin of the vehicle are highly comparable to the results obtained based on vehicle value. This suggests that wealth affect is approximated to that of vehicle value.

In a paper published in the Federal Reserve Bank of Chicago Economic Perspective in 2008, Agarwal et al. (2008) discuss the determinants of automobile loan default and prepayment. They build a competing risks framework to analyze the prepayment and default options on auto loans. They use data from a large financial institution that originates direct automobile loans. They study loan characteristics such as automobile value, automobile age, loan amount, LTV (loan to value), monthly payments, contract rate, time of origination (year and month), as well as automobile make, model, and manufacturing year. After analysis, they conclude that loans on new cars have a higher possibility of repayment while loans on used cars have a higher possibility of default. Moreover, the higher the credit score of loan holder, the higher the possibility of repayment. The higher the automobile worth, the higher the possibility that the loans would default. An increase in income raises the probability of prepayment, whereas a rise in unemployment increases the probability of default. The most interesting finding is that luxury cars have higher possibility of prepayment while most economy automobiles have a lower probability of default.

The same authors have also studied the impact of asymmetric information on the automobile loan market. They use a unique dataset of individual automobile loans to assess whether borrower consumption choice reveals information about future loan performance. In order to find out the answer, they adopt a competing-

risks framework to analyze auto loan prepayment and default risks. The findings can be summarized in the following four points: first of all, a decrease in borrower's credit risk leads to higher prepayment rate and lower default rate; secondly, an increase in the loan-to-value ratio increases the risk of default and lowers the likelihood of prepayment; thirdly, a decrease of borrower income or increase in local unemployment rate increases the default rate; lastly, a decrease in the market interest rate increases both the probability of prepayment and default. More interestingly, they also find that vehicle manufacturer location (America, Europe, and Japan) significantly impacts both the repayment and default behavior of borrowers.

The research done by Ambrose and Sanders (2003) on the prepayment and default of commercial mortgage-backed securities suggests that changes in the yield curve have a direct impact on the probability of mortgage termination. However, they found no evidence suggesting that LTV has a significant relationship with prepayment of default.

Heitfield and Sabarwal (2004) conduct the only other study on default and prepayment for automobile loans. They use performance data from subprime auto loan pools underlying asset backed securities. Through analyzing these underlying loan contract performances, they conclude that prepayment rates increase rapidly with loan age but are not affected by prevailing market interest rates. Moreover, their data also support the idea that increases in unemployment precede increases in default rates, suggesting that house liquidity has a huge impact on automobile loan performance.

By studying a large auto finance company and the changes after it adopted lending practices, Einav et al. (2013) suggest that credit scoring appears to have increased profits by roughly a thousand dollars per loan. They have also identified two direct benefits of credit scoring: the ability to screen high risk clients as well as the ability to target more generous loans to lower-risk borrowers.

Little research has been done on the online lending sector. One paper which is relevant focuses on predicting prepayment and default risks of unsecured consumer loans in online lending. Li et al. (2018) found that online lending contracts have high default risk compared to that of normal off-line loans. They also have higher proportions of early repayment. High interest rates charged on a loan do not only indicate high probability of default but also increase the probability of prepayment. The borrower characteristics such as the debt-to-income ratio and credit score have significant impact on both outcomes. Macroeconomic factors such as GDP growth, the Federal fund rates and the personal bankruptcy rate can also influence the occurrence of the two events.

Lin et al (2017) have done research on the credit risk of online Peer-to-Peer lending businesses by exploring the factors that determine default risk based on the demographic characteristics of borrowers. The empirical data were generated from a large P2P platform in China. In order to quantify the default risk of each loan,

they propose a credit risk evaluation model by estimating a binary logistic regression. They suggest that borrowers with low default risk fit into the following: young adults; females; people with a long work history, in stable marriages, enjoying high levels of education, and employed by large companies; consumers making low monthly payments, carrying small loan amounts, benefiting from a low debt-to-income ratio and having no default history.

Zhou, Zhang & Luo. (2018) have conducted research on P2P network lending and focused on loss given default (LGD) and credit risks. Their paper suggests that the distribution of LGDs of P2P lending is similar to that of unsecured bonds. They also discover that the total loan volume has little impact on loss given default; however, the credit rating and debt-to-income ratio are strongly dictated by the default event.

Polena & Regner, Tobias. (2018) suggest that the debt-to-income ratio, inquiries in the past six months, and loans intended for small business are positively correlated with the default rate. Annual income and credit card as loan purpose are negatively correlated. The data they use are from a new data set consisting of 70,673 P2P loan observations from the Lending Club.

Online lending is quite different from traditional lending in terms of data gathering and data analysis. By collecting client mobile usage data such as series of records from the call, message, data volume, and App usage, Liu, Ma, Zhao & Zou (2018) could describe user behaviours and propose a model that can accurately predict the default rate. They found that personality traits, socioeconomic status, consumption patterns, and economic characteristics are correlated with credit default behaviour. This model has been estimated by using real life data and has yielded satisfactory performance.

Some research produces results that are contradictory to this literature. We note that, compared with divorced clients, (other variables controlled), single clients have higher default probability than that of married clients. However, Suzuki. (2018) has conducted research on Latin American auto loan default. The paper suggests that debtors who got married and had a family tended to default more as their expenditure is larger than unmarried or single people. Moreover, when the information of down payment and income is missing, the default rate increases. This suggests that risk factors are very different depending on the disclose of information.

Gender discrimination has long been discussed in the lending business and the credit score rating method. For example, Apple credit card has been criticized for giving higher credit amounts to men than to women by using its own credit ration algorithm. However, Lin. (2019) suggests that there is no significant gender effect on the probability of default, ceteris paribus. This result is derived from analyzing the data from a big P2P online lending platform. Therefore, it can be said that a borrower's gender is not good screening

criteria to control the credit risk. The author also mentioned that this conclusion might be only platform specific and it should not be simply applied to other platforms without a test.

IV. **Data description**

The data used in the analytical part of this work come from ASFC, a division of Peak Positioning Inc., a financing institution, whose operations include direct personal loans. Sample data come from clients and are generated from the auto loan business. Client information is collected at the time when they sign an auto loan contract as it is obligatory for them to provide personal information (identity cards, etc.)

The goal of this data analysis is to evaluate the importance of each factor which affects the fact that clients will default or will make a prepayment on their automobile loan. The goal of variable classification is to effectively predict at ex-ante default and prepayment events in order to enhance the efficiency of the early warning risk management system.

This original raw dataset is cross-sectional and includes 19 variables and 4557 records within which the earliest loan contract started on 2014-10-11 and the latest loan contract ended on 2019-08-01. This data set also contains several contract terminations between 2020 and 2021, as loans have already closed due to the prepayment. Ultimately, we consider all these entries as valid prepaid loans.

The cleaning of dataset was also performed to eliminate the entries that have missing variables, duplicated information, incomplete or on-going loan contracts and confusing or wrong data input. This data-purification step decreased the size of data sample to 3665 entries which represents 80% of original data size.

Since we are interested in the behaviour of clients who ended up in default or repaid the loan before the contract matured, our original data set have three selected groups of 19 variables which cover the customers' personal information, automobile data, and the information of loan contracts:

1) Customer data are their social demographic and demographic characteristics including residence city on ID, birthday, gender, marital status;
2) Automobile data such as auto models, auto net value, vehicle license registration location, nature of ownership (private or corporate);
3) Loan contract information consists of loan amount, loan status (default or well closed), loan form (support by credit or automobile collateral immobilized in institutional lender), LTV rate (loan amount / collateral value), repeated borrowing date (no values means it is a one-time borrowing), loan actual closing date, loan contract beginning date, loan contract ending date, payment type (equal total loan payment or interest first payment), loans maturity periods (3, 6, 12, 24 months), and monthly payment amount.

In order to maximize the use and the insights of data information, based on the above-mentioned 19 variables, we have also reorganised or transformed some variables:

Default, prepayment: One of the dependent variables in this analysis, *default* (= 1 if default event occurred or beyond one-week delayed payments from loan's contractual mutuality date observed, = 0 if not), is directly obtained from the original database. The institution identifies an individual to be a defaulter whenever there are no payments for more than a two weeks' delay[2]. Another dependent variable, *prepayment*, is calculated from the original database (= 1 if auto loan's actual closing date is at least 2 months earlier than the loan's contractual ending date.) These two dependent variables are instinctively excluded from each other, hence they practically serve as a competing risk model in order to mutually verify significance and influence of explanatory variables.

Age: As we focus on clients' personal status when he or she decided to prepay or fail a loan, we calculate the difference between clients' birthday and auto loan's actual closing date as clients' age variable.

Clear date / Beginning date / Ending date: in the original dataset, *clear date* is auto loan's actual closing date, *beginning date* is auto loan's contractual beginning date, similarly, the *ending date* is the contractual ending date. As all these three variables are in date data type, for better modeling purposes, we later transform them in continuous data type. They are replaced by calculating the number of days from the eve of the first beginning date (which is considered the opened date of auto loan business in institution: 2014-10-15) to the three corresponding date variables. Furthermore, with the assumption that the default and prepayment event may have seasonal characteristics, Agarwal et al. (2008) include quarter dummies in the auto loan competing risks model. However, considering the time span in our dataset is too long (from 2014 to 2021), we only include annual dummies corresponding to clear year, beginning year, and ending year.

Repeatedly: repeated borrowing (= 0 means one time borrowing, = 1 means the repeated borrowing with or renewed loan contract once, = 2 twice, etc.) As the dataset is cross-sectional, the same client with the same or different automobile collaterals could obtain different loan contracts in serial timeline. In the original dataset, this type of client is recorded as a new client each time when they write down a new or renewed loan contract. However, some personal characteristics of these clients remain unchanged (ex: information obtained from identity card). Furthermore, as we assume that the repeated borrowing action may be connected to the default or prepayment event, we still consider this type of record as a new client's record. However, this situation is controlled by the variable of the repeated borrowing number.

Product: inspired by the analysis of Dionne & C. (1993) about the presence of adverse selection in the automobile insurance market, we assume that different loan products chosen by clients reveal their risk exposure: clients choose different reimbursable methods with the consideration of their termly payment's

[2] Given that auto loans in this study are mostly short-term loans for three months, according to the recommendation of the director of the risk management department of ASFC, if no payment is received for more than a two weeks' delay, the risk management department will begin to take necessary actions, such as taking control of the automobile or starting legal proceedings.

ability. Variable *Product* connects two kinds of loan reimbursable methods with different maturity terms offered by institution. Different maturities offered are 3, 6, 12, 24 months, and two kinds of loans payment offered are:

- *Equal total loan payment*: all the principal and interest are generated and evenly distributed every month, meaning that the monthly reimbursement is the same across time. The amount applied towards the principal increases with each payment. It is suitable for large consumption, especially consumption beyond short-term payment capacity, usually with relatively long-term maturity.
- *Interests payment first and then principal paid in full at maturity*: it refers to one-year consumer credit. This repayment method is more suitable for clients suffering from short-term cash flow pressure.

In the *product* variable, the value *I3* indicates that interest is paid by the clients at first and then the principal is paid in full at the end of 3-month maturity. In a similar way, *IP12* means clients choose equal total loan payment with maturity at 12 months.

Duration: number of months from loan contractual beginning date to loan actual closing date. As clients may prepay the auto loan, this variable is different from the loans' contractual maturity, which explains the prepayment behaviour.

Licence: In the original raw dataset, there are 15 vehicle license registration locations, among which the location *Wuxi* (the city where the financing institution is located) has been recorded 3603 times within 3665 total data records. We thus reorganized this variable as a dummy variable (= 1 if automobile's license is issued at *Wuxi*, = 0 otherwise.)

Region (automobile manufactured location): We could observe 668 automobile models recorded in the original dataset. We need to reduce the categories to do a better modelling, so we reorganised the model variables in restricted 64 marks categories, then allocated to 6 automobile manufacturer locations: *EU (Europe), Japan, China, Korea, USA, UK*.

Province (clients' residential province): The original dataset recorded 382 clients' residential cities noted in their identity card, and we regrouped the 382 cities in 23 clients' residential provinces. As 23 provinces is still abundant for a categorical variable, we transformed it in dummy variable (= 1 if clients' residential provinc*e* is *Jiangsu* (the province where the financing institution is located an business covered), =0 otherwise.)

We also assumed that default event relates to social economic environment. More default events are observed usually in recession. Therefore, we also included a group of 3 variables indicating economic aspect as the fourth group variables in addition to the above-mentioned 1), 2), and 3) group variables: 4)

Economic variables: *sales* (retail sales of consumer goods), *gdp* (GDP indicators), *income* (the per capita disposable income of provincial residents).

This group of variables were collected from the Statistics Bureau of Jiangsu province in China[3]. We only considered the data of Jiangsu province here because this is the region mainly covered by the financing company's business and clients base. We controlled unobserved heterogeneity by the *province dummy* mentioned above for clients whose residence is not in Jiangsu. As available data is seasonal, for each record these three variables are calculated as the monthly average in the length of auto loan's contractual beginning date to the actual ending date time window. Seasonal data are converted to monthly data by dividing by three. Then if the contractual beginning date does not start at the first day of a month, we calculate the daily average and multiply the previous number by the number of contractual days in this month. All these three variables are in units of trillion CNY.

In the end, considering that almost half of the variables in the original dataset are recorded in character form, we transform all the binomial variables in dummy form: =1 means the observation is in the corresponding categories, = 0 otherwise. For example, *Loan form dummy, Gender dummy, License dummy, Nature dummy, Province dummy.* Categorial variables are also transformed to dummies: *Marriage (_married, _single) dummy, Repeatedly (_0,_1_2,_3,_4) dummy, Product(_I3, _I6, _I12, _PI6, _PI12, PI_18, PI_24) dummy, Clear year(_2015,_2016, _2017, _2018, _2019) dummy, Beginning year(_2014,_2015,_2016, _2017, _2018, _2019) dummy, Ending year(_2015, _2016, _2017, _2018, _2019, _2020) dummy.*

Variables used in this analysis are presented in Table 1 below.

Table 1. List of variables used in analysis

Variable		Description
Dependent variable		
Default	=	1 if default happened, 0 if no default
Prepayment	=	1 if client prepaid, 0 otherwise
Independent variable		
Quantitative variable		
Loan (thousand CNY)	=	auto loan amount attributed to client
LTV (loan to value)	=	auto loan amount / automobile value
Age	=	age of client at the loan actual closing date

[3] Source: http://tj.jiangsu.gov.cn/col/col72292/index.html

Table 1. (continued)

Clear date	=	loan actual closing date
Beginning date	=	loan contract beginning date
Ending date	=	loan contract ending date
Duration (month)	=	number of months from loan contractual beginning date to loan actual closing date
Monthly payment (thousand CNY)	=	monthly payments for auto loan
Car value (thousand CNY)	=	value of the collateral automobile
Income (thousand CNY)	=	monthly average of per capita disposable income of provincial residents
Sale (trillion CNY)	=	monthly average of provincial retail sales of consumer goods
GDP (trillion CNY)	=	provincial GDP indicator
Dummy variable		
Loan form	=	1 if loan is backed by client's credit, 0 if auto is collateralized immobilized in intuitional lender
Gender	=	1 if client is female, 0 otherwise
License	=	1 If auto is registered in Wuxi, 0 otherwise
Nature	=	1 if auto is a personal property, 0 is cooperate
Province	=	1 if client is a native resident, 0 otherwise
Categorial dummy variable		
Marriage (_married, _single)	=	1 if in this category, 0 otherwise. (_divorced is omitted here as referential category)
Repeatedly (_0,_1_2,_3,_4)	=	1 if in this category, 0 otherwise. Repeatedly time when client reborrow or renew the contract of auto loan (_4 is omitted here as referential category)
Product (_I3, _I6, _I12, _PI6, _PI12, PI_18, PI_24)	=	1 if in this category, 0 otherwise. Various loans' reimbursable methods with different maturity offered by institution (_PI3 is omitted here as referential category)
Clear year (_2015,_2016, _2017, _2018, _2019)	=	1 if in this category, 0 otherwise. Loan actual closing year (_2015 is omitted here as referential category)
Beginning year (_2014,_2015,_2016, _2017, _2018, _2019)	=	1 if in this category, 0 otherwise. Loan contractual beginning year (_2014 is omitted here as referential category)
Ending year (_2015, _2016, _2017, _2018, _2019, _2020)	=	1 if in this category, 0 otherwise. Loan's actual ending year (_2015 is omitted here as referential category)

1. Univariate statistics

Table 2 to Table 4 show the summary of univariate statistics. Table 2 shows outcomes and frequencies for dependent variables, default and prepayment, respectively. We observe 297 default events which represents 8.1% of the total 3365 clients base. As institutions only accept loan requests after some basic assessment of client risk exposure, a client's loan request will be refused immediately if the client is deemed to be risky. This high percent of default event in database could be explained by the defective existence of ex-ante risk

control. For prepayment variable, 1456 clients representing 39.73% of total data chose to prepay their auto loan. Prepayment event is defined as client paying in full their auto loan more than two months in advance to loan's contractual maturity date. Figure 3 presents histograms of these two variables, which shows especially for default event that the original database is imbalanced. This imbalance could affect data training. (We will consider this problem in a following section.) What's more, both variables have two outcomes: 0 and 1. Therefore, we conclude that data is convenient for discrete choice modeling. This gives us an idea to build binary Probit or Logit models, which are commonly used in related analysis and research.

Figure 3. Dependent variables – Histograms chart

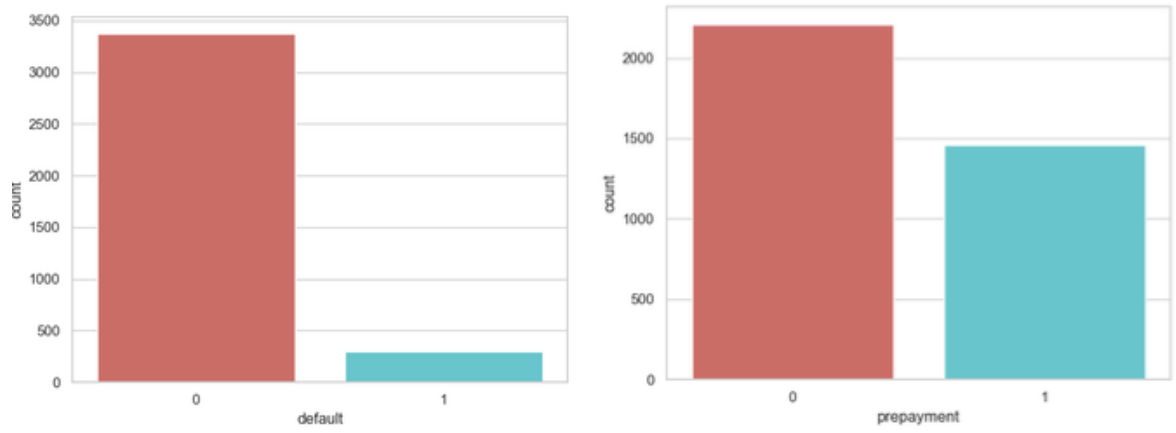


Table 2. Univariate statistics – Dependent variables

Dependent variable	Outcome	Frequency	Percent
Default	0	3368	91.9
	1	297	8.1
Prepayment	0	2209	60.27
	1	1456	39.73

Table 3 gives the 25th percentile, mean, 75th percentile and the standard deviation of the quantitative variables used. Most of the company's car loan business was transacted between 2016 to 2018. The average loan amount is 93.725 thousand CNY with a 91.259 standard deviation which in turn implies a dispersion in the distribution of loan amount. The mean of *LTV* (load amount / auto's market value) is 0.812. Most auto loans have a short-term duration. The client age falls mainly into the 30-to-45-year range . The car

market value in the institution ranges from 50 thousand to 138 thousand CNY. These auto loan statistics are comparable with the overall statistics of industry.

Table 3. Univariate statistics – Quantitative independent variables

Quantitative variable	25th percentile	Mean	75th percentile	Std
Loan (thousand CNY)	40.000	93.725	110.000	91.259
LTV (auto loan/auto value)	0.796	0.812	0.8966	0.150
Age	30.000	37.234	44.000	9.312
Clear date	20160721[4]	20170618	20180510	400.519
Beginning date	20160106	20161114	20170820	373.565
Ending date	20161007	20170909	20180720	416.993
Duration (month)	3.000	7.144	12.000	4.662
Monthly payment (thousand CNY)	2.625	8.747	8.099	24.833
Car value (thousand CNY)	50.000	117.630	138.000	116.476
Income (thousand CNY)	2.578	3.167	3.144	1.373
Sale (trillion CNY)	0.235	0.287	0.2769	0.124
GDP (trillion CNY)	0.592	0.721	0.7249	0.303

Table 4. Univariate statistics – Categorial independent variables

Categorial variable	Outcome	Frequency	Percent
Loan form	credit	3538	96.53
	collateral immobilized	127	3.47
Gender	M	2919	79.65
	F	746	20.35
Marriage	married	2984	81.42
	single	474	12.93
	divorced	207	5.65
Repeatedly	0	2119	57.82
	1	1222	33.34
	2	247	6.74
	3	59	1.61
	4	17	0.46
	5 (dropped)	1	0.03

[4] As mentioned before, variables *clear date*, *beginning date*, *ending date* are transformed from date type data to continuous type data; the *percentiles*, *mean,* and *standard deviation* calculated here were number of days. For a better understanding, we transformed the *percentiles* and *mean* back to date type data by adding corresponding number of days calculated after the eve of the first beginning date (which is considered the opened date of auto loan business in institution: 2014-10-15) and get the date type data represented in the table.

Table 4. (continued)

Product	I3	1138	31.05
	I6	39	1.06
	I12	16	0.44
	PI3	8	0.22
	PI6	142	3.87
	PI12	1799	49.09
	PI18	435	11.87
Licence	wuxi	3603	98.31
	ex_wuxi	62	1.69
Nature	private	3398	92.71
	cooperate	267	7.29
Region	EU	1221	33.32
	Japan	1067	29.11
	USA	635	17.33
	China	484	13.21
	Korea	176	4.8
	UK	82	2.24
Province	native	3046	83.11
	ex	619	16.89
Clear year	2015	412	11.24
	2016	935	25.51
	2017	1033	28.19
	2018	951	25.95
	2019	334	9.11
Ending year	2014 (dropped)	1	0.03
	2015	245	6.68
	2016	892	24.34
	2017	989	26.98
	2018	1031	28.13
	2019	430	11.73
	2020	75	2.05
	2021 (dropped)	2	0.05
Beginning year	2014	32	0.87
	2015	869	23.71
	2016	1039	28.35
	2017	1137	31.02
	2018	546	14.90
	2019	42	1.15

Table 4 gives statistics for categorial independent variables used. These numbers indicate that only 3.47% of autos judged as higher risk are collaterals immobilized. Most clients are male, married, and native resident; they reborrow or renew auto loan contracts for no more than 2 times, and prefer to pay the 3-month interest payment first and then the principal paid in full at maturity or one year's equal total loan payment loan product. Up to 90 percent of cars are private property with a native license and 80% of them are manufactured in Asia or Europe. Finally, we drop from dataset the categories with only one or two

observations, which reduce the number of *ending year dummies* to 6, the number of *clear year dummies* to 5.

2. Bivariate statistics

Tables 5 and 6 present the bivariate statistics of dependent variables vs independent variables. Mean and standard deviation of independent variables are calculated separately for defaulter and non-defaulter, client prepaid and non-prepaid datasets. T-statistics and their correspondent P-value are provided to test if these two separate datasets have the same distribution. We only present in Table 5 and 6 the variables which significantly reject H0. Stake Bar chart of selected variables are provided for a better intuitive understanding of the dataset.

Table 5. Bivariate statistics – Default vs independent variables

***Default* =**		**Mean**	**SD**	**T-statistic**	**P-value**
Prepayment	0	0.432	0.495	-15.0348	0.000
	1	0.000	0.000		
Duration(month)	0	6.936	4.531	9.209	0.000
	1	9.505	5.418		
Income(thousand CNY)	0	3.188	1.410	-3.139	0.002
	1	2.928	0.811		
Sale(trillion CNY)	0	0.289	0.128	-3.176	0.002
	1	0.265	0.065		
GDP(trillion CNY)	0	0.725	0.313	-2.432	0.015
	1	0.680	0.158		
Clear date	0	20170613[5]	400.768	2.251	0.024
	1	20170807	394.897		
Region_Korea	0	0.045	0.208	2.759	0.006
	1	0.081	0.273		
Clear year_2016	0	0.262	0.440	-3.305	0.001
	1	0.175	0.381		
Clear year_2017	0	0.28652	0.452	-2.114	0.035
	1	0.228956	0.421		
Clear year_2018	0	0.249	0.433	4.699	0.000
	1	0.374	0.485		
Ending year_2014	0	0.000	0.000	3.372	0.001
	1	0.003	0.058		

[5] As mentioned before, variables *clear date, beginning date*, *ending date* are transformed from date type data to continuous type data, the *mean* and *standard deviation* calculated here were number of days. For a better understanding, we transformed the *mean* back to date type data by adding corresponding number of days calculated after the eve of the first beginning date (which is considered the opened date of auto loan business in institution: 2014-10-15) and get the date type data represented in the table.

In default and non-default datasets, the average duration of default loan is longer than that of non-default loan. All three economic variables are significant, which confirms our assumption that there are fewer defaulters when the economy is in a good state. There are fewer defaulters if the auto loan is closed in 2016 and 2017, and conversely in 2018, the percentage defaulters increased. What's more, clients who possess a car made in Korea have a higher default probability. We can also observe these phenomena from the stake bar chart below. In the bar chart, clients who have a car collaterally immobilized in institutional lenders seem to have a higher default probability. Institutions may already judge such clients as having high risk exposure. In addition, marital status of clients seems to be also a good predictor for no default.

Figure 4. Selected variables VS default - Stake Bar chart

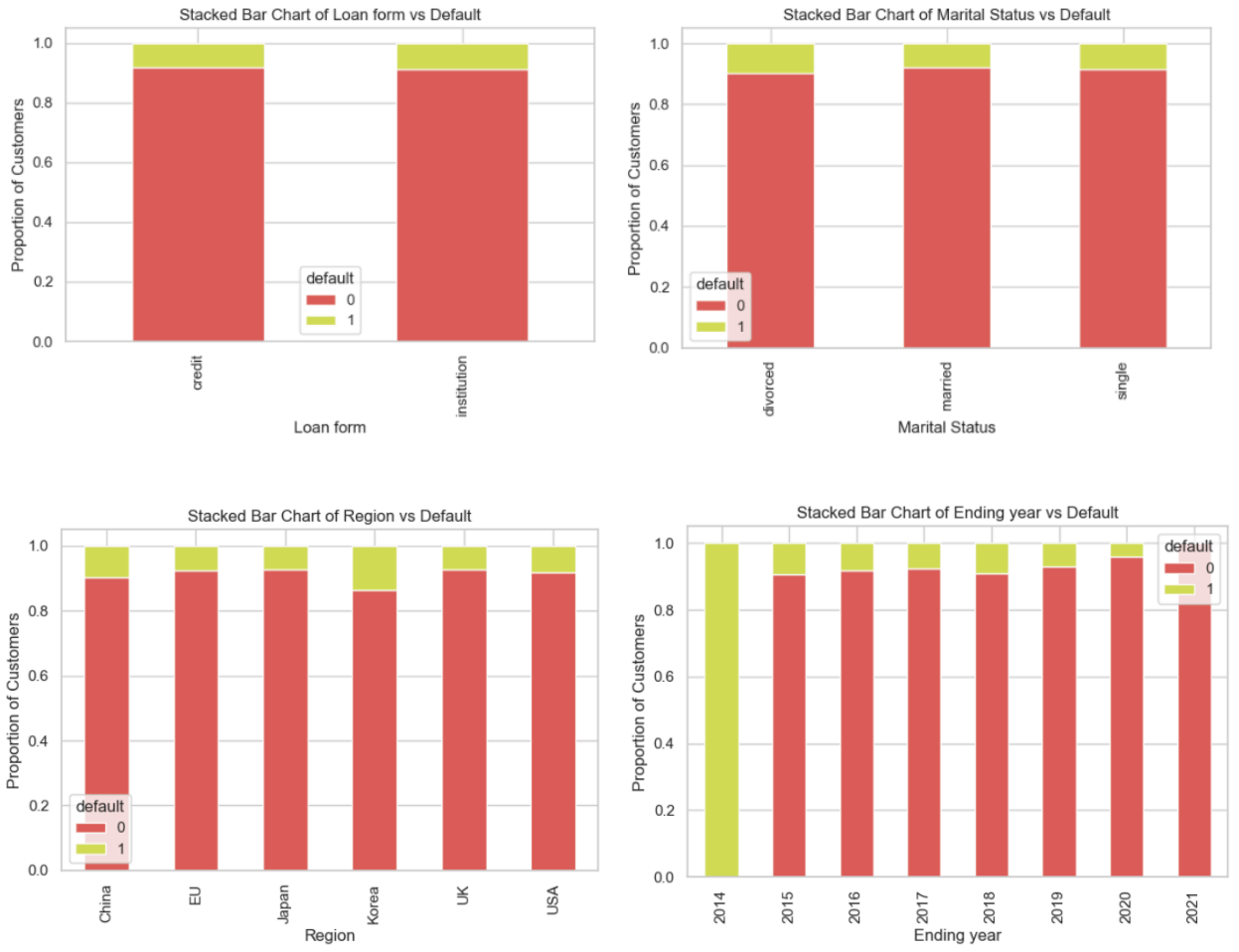


Table 6. Bivariate statistics – Prepayment vs independent variables

Prepayment =		Mean	SD	T-statistic	P-value
Default	0	0.134	0.341	-15.035	0.000
	1	0.000	0.000		
Age	0	37.993	9.495	-6.105	0.000
	1	36.083	8.908		
Repeatedly	0	0.599	0.772	-6.414	0.000
	1	0.441	0.666		
Duration (month)	0	7.575	5.223	-6.948	0.000
	1	6.489	3.553		
Car value (thousand CNY)	0	120.702	130.130	-1.968	0.049
	1	112.968	91.816		
Income (trillion CNY)	0	3.055	1.101	6.112	0.000
	1	3.337	1.690		
Sale (trillion CNY)	0	0.276	0.099	6.442	0.000
	1	0.303	0.153		
GDP (trillion CNY)	0	0.694	0.247	6.678	0.000
	1	0.762	0.369		
Clear date	0	20170721[6]	406.801	-6.224	0.000
	1	20170428	385.489		
Beginning date	0	20161204	379.152	-4.014	0.000
	1	20161015	362.941		
Ending date	0	20170725	407.880	8.384	0.000
	1	20171119	420.911		
Loan form_credit	0	0.956	0.206	3.967	0.000
	1	0.980	0.140		
Gender_F	0	0.223	0.416	-3.641	0.000
	1	0.174	0.379		
Marriage_married	0	0.831	0.375	-3.168	0.002
	1	0.789	0.408		
Marriage_single	0	0.115	0.319	3.293	0.001
	1	0.152	0.359		
Repeatedly_0	0	0.537	0.499	6.264	0.000
	1	0.641	0.480		
Repeatedly_1	0	0.361	0.480	-4.412	0.000
	1	0.291	0.454		
Repeatedly_2	0	0.074	0.262	-2.037	0.042
	1	0.057	0.232		
Repeatedly_3	0	0.021	0.144	-3.071	0.002

[6] As mentioned before, variables *clear date, beginning date*, *ending date* are transformed from date type data to continuous type data, the *mean* and *standard deviation* calculated here were number of days. For a better understanding, we transformed the *mean* back to date type data by adding corresponding number of days calculated after the eve of the first beginning date (which is considered the opened date of auto loan business in institution: 2014-10-15) and get the date type data represented in the table.

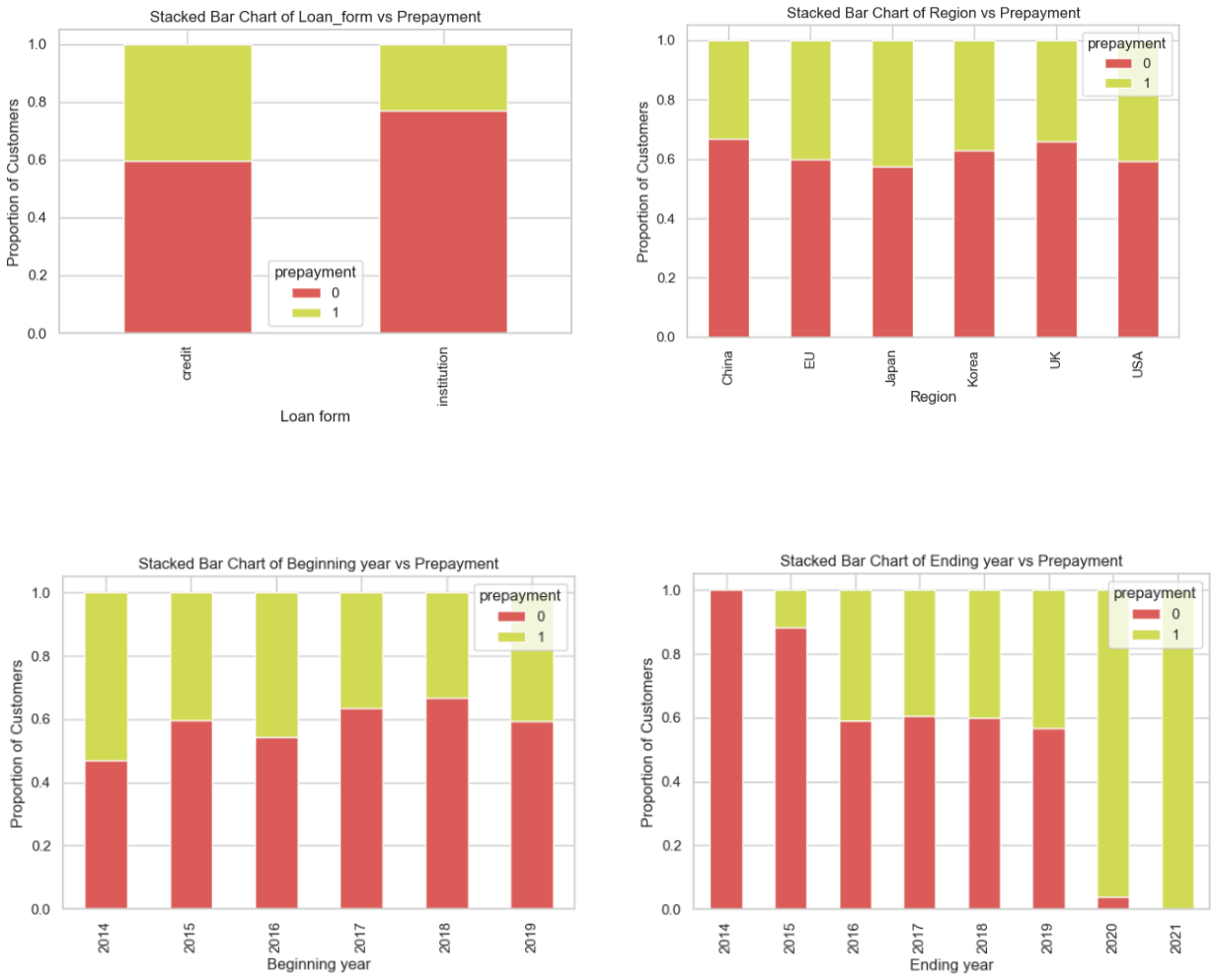
Table 6. (continued)

	1	0.008	0.090		
Product_I3	0	0.480	0.500	-30.567	0.000
	1	0.054	0.225		
Product_PI12	0	0.373	0.484	18.441	0.000
	1	0.670	0.470		
Product_PI18	0	0.077	0.267	9.635	0.000
	1	0.181	0.385		
Product_PI24	0	0.002	0.043	10.988	0.000
	1	0.058	0.233		
Product_PI3	0	0.004	0.060	-2.300	0.022
	1	0.000	0.000		
Product_PI6	0	0.051	0.219	-4.632	0.000
	1	0.021	0.142		
Region_China	0	0.146	0.353	-3.122	0.002
	1	0.111	0.314		
Region_Japan	0	0.278	0.448	2.238	0.025
	1	0.312	0.463		
Nature_private	0	0.917	0.276	3.000	0.003
	1	0.943	0.232		
Province_native	0	0.822	0.383	1.794	0.073
	1	0.845	0.362		
End year_2015	0	0.098	0.297	-9.342	0.000
	1	0.020	0.140		
End year_2020	0	0.001	0.037	10.202	0.000
	1	0.049	0.217		
End year_2021	0	0.000	0.000	1.743	0.081
	1	0.001	0.037		
Clear year_2016	0	0.229	0.420	4.546	0.000
	1	0.295	0.456		
Clear year_2018	0	0.285	0.451	-4.307	0.000
	1	0.221	0.415		
Clear year_2019	0	0.108	0.310	-4.313	0.000
	1	0.066	0.248		
Beginning year_2016	0	0.255	0.436	4.674	0.000
	1	0.326	0.469		
Beginning year_2017	0	0.327	0.469	-2.680	0.007
	1	0.285	0.452		
Beginning year_2018	0	0.165	0.371	-3.314	0.001
	1	0.125	0.331		

In prepayment and non-prepayment datasets, male clients are more likely to prepay than female clients. Clients who choose the equal total loan payment with more than one year's maturity loan product have

more propensity to prepay. Single clients seem to have higher probability to prepay than that of married clients. An increase in number of reborrowing or renewal of loan contract by customer, yields a higher possibility that they will prepay. Native clients opt more for prepayment than non-native clients. The stake bars below also tell us that clients who have a car that is collateral immobilized in institutional lender have less probability to prepay, and clients who have an economical car (lower *car value)* or made in Japan have tendency to prepay. According to Dionne &Lui (2017), if we regard car value as a proxy of a client's wealth, this may refer to a wealth or income effect on prepayment probability. We could observe many variations in *beginning year* and *ending year* variables from 2014 to 2019, which implies that prepayment motivation may depend on the time state.

Figure 5. Selected variables VS prepayment - Stake Bar char



V. Methodology

As a first step, we randomly split datasets into training data and test data by a ratio of 7 : 3. Training datasets serve to training models and test models will serve to examine the predictive performance of models. Furthermore, considering that our dataset classes are imbalanced, the ratio of defaulter to no-defaulter is 8.1 : 91.9. Before the implantation of models, it is necessary to balance the classes, with our training data created. We up-sample the defaulters using the SMOTE algorithm (Synthetic Minority Oversampling Technique.) At a high level, instead of creating copies, SMOTE works by creating synthetic samples from the minor class defaulters. This model randomly chooses one of the k-nearest-neighbors and uses it to create similar, but randomly, new observations[7].

Table 7. Data Oversampling

	Default Model				**Prepayment Model**			
	Before oversampling		After oversampling		Before oversampling		After oversampling	
Data length	3365		4722		3365		3066	
	No default	Default	No default	Default	No prepay.	Prepay.	No prepay.	Prepay.
Frequency	3368	297	2361	2361	2290	1456	1533	1533
Percent	91.9%	8.1%	50%	50%	60.27%	39.73%	50%	50%

Before oversampling, the percentage of non-default is 91.9 and the percentage of default is 8.1. After oversampling, the length of oversampled data becomes 4722. The number of non-default and defaults are 2366 each. The proportion of non-defaulter and defaulter in the oversampled data is balanced to 1 : 1.

Before oversampling, the percentage of non-prepayment is 60.27 and the percentage of prepayment is 39.73. After oversampling, the number of oversampled data is 3066 and the number of non-prepayment and prepayment are 1533 each. The proportion of non-prepayment and prepayment in oversampled data is again balanced to 1 : 1.

We need to point out that we over-sampled only on the training data, because by oversampling only on the training data, none of the information in the test data is used to create synthetic observations. Therefore, no information will bleed from test data into the model training.

[7] Source: https://imbalanced-learn.readthedocs.io/en/stable/generated/imblearn.over_sampling.SMOTE.html

1. Implementing Models

Since our primary purpose is to extract determinant factors that affect auto loan performance, we follow the methodology used by Agarwal et al. (2008) to estimate competing-risks models of auto loan default and prepayment. It is called "competing" because of that clients who have defaulted and show absolutely no possibility of prepaying. In each modeling method, default and prepayment will be dependent variable side by side. Also following Agarwal et al. (2008) and Gross and Souleles (2002), we group X_i into components representing loan borrower personal characteristics, collateral car characteristics, signed auto loan product characteristics, local economic conditions and series of dummies corresponding to calendar years that allow shift overtime.

In high level, we assume that :

$$X'\beta = \beta_0 + \beta_1 year\ dummy_{i1} + \beta_2 borrower_{j2} + \beta_3 car_{k3} + \beta_4 loan_{l4} + +\beta_5 economic_{m5} \qquad (1)$$

Where $year\ dummy_i$ represents a series of dummies corresponding to calendar year (*beginning year dummies, clear year dummies, ending year dummies*); $borrower_j$ represents a set of continual or dummy variables of client's social-geographies characters (*age, gender dummy, marriage dummy, residential province dummy*); car_k represents a set of continual or dummy variables concerning collateral cars information (*car value, licence dummy, nature dummy, manufacturer region dummies*); $loan_l$ is a set of variables concerning the auto loan product taken by client (*loan, LTV, duration, monthly pay, clear date, beginning date, ending date, loan form dummy, repeatedly dummies, product dummies*); and $economic_m$ is a set of variables capturing local provincial economic conditions (*sales, gdp, income.*)

Equation (1) contains all the basic variables some of which are widely used in the analysis of personal loan performance. We then scale down the tribe of independent variables to pick only the variables which are the most robust and have the most explanatory power. By comparing the log-likelihood ratio statistics of the model, we can then determine the marginal impact of integrating information parameter above into the assessment of the likelihood of default or prepayment repayment.

Since the two dependent variables in the competing-risks model (*default* and *prepayment)* are binary variables with value 0 or 1, for modeling purposes, we shall first consider discrete choice models. The populated linear probability model (OLS) has a major weakness because it assumes the conditional probability function to be linear. It has the clear drawback of not being able to capture the nonlinear nature of the population regression function and it may not restrict predictable probabilities $P(Y = 1|X_1, X_2, .. X_k)$ to lie inside the interval [0,1], beyond which the model has no meaningful interpretation. This inconvenience calls for an approach that uses a nonlinear function to model the conditional probability function of a binary dependent variable. Commonly used methods are Probit and Logit

regression. The Probit model and the Logit model deliver only approximations to the unknown population regression function $E(Y|X)$. Therefore, they are harder to interpret but they capture nonlinearities better than the linear approach. Both models produce predictions of probabilities that lie inside the interval [0,1]. The major difference between Probit and Logit is the distribution of $X'\beta$. The Probit assumes $X'\beta$ has a normal distribution, while Logit assumes $X'\beta$ has an exponential distribution. But if the number of samples exceeds 500, the differences in the overall results of the models are usually minimal and almost none, which means that it is still ambiguous to decide which model is more appropriate in practice.

Model 1. Probit Regression

In the Probit regression, when the dependent variable Y is binary, the cumulative standard normal distribution function $\Phi(\cdot)$ is used to model the regression function, that is, we assume:

$$E(Y|X_1, X_2, ..X_k) = P(Y = 1|X_1, X_2, ..X_k) = \Phi(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k)$$

$X_1, X_2, ..X_k$ are multiple explanatory regressors. $(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k)$ plays the role of a quantile z. Remember that $\Phi(z) = P(Z \leq z), Z \sim N(0,1)$.

Although the effect on z of a change in X is linear, the link between X and the dependent variable Y is nonlinear since Φ is a nonlinear function of X. The Probit coefficient β_i is the change in z associated with a one unit change in X_i, holding constant all other regressors.

In Python, Probit models can be estimated using the *spreg.Probit*.[8]

Model 2. Logit Regression

The Logit regression function is:

$$P(Y = 1|X_1, X_2, ..X_k) = F(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_k)}}$$

The idea is similar to Probit regression except that a different CDF is used: $F(X) = \frac{1}{1+e^{-(x)}}$

As we have mentioned above, in practice, the predictions made by the two models discussed above are close to each other, and therefore give no general recommendation as to which method to use. In this analysis, we conclude later that Logit model has a slightly better performance than Probit model, but this difference is not significant.

[8] Source : https://pysal.readthedocs.io/en/v1.11.0/library/spreg/Probit.html

In Python, Logit models can be estimated using the *Logit module* in statsmodels.api package.[9]

Model 3. Logit with recursive feature elimination.

Here we use a feature ranking system with RFE (recursive feature elimination) in order to select out the most important determinants. According to Guyon et al. (2002), RFE is basically the backward selection of explanatory variables. This technique first builds a model on the entire set of predictors and then calculates an importance score for each predictor. Next, it deletes the least important predictor, rebuilds the model, and calculates the importance score again. In effect, the analyst specifies the number of prediction subsets to be evaluated and the size of each subset. Therefore, the subset size is a tuning parameter of the RFE. The subset size of the optimized performance criteria is used to select predictors based on importance rankings. The best subset is then used to train the final model.

As mentioned earlier, based on a basic model, the Logit model in our case, the goal of recursive feature elimination (RFE) is to select features by recursively considering smaller and smaller sets of features. In Python, the estimator is trained firstly on the initial set of features, and the importance of each feature is obtained either through a *coef* attribute or through a *feature importance* attribute. The *feature importance* attribute is calculated according the positive change of log-likelihood of the model (Logistic model in our case) by adding and removing consecutively each feature. Then, after the *feature importance* attribute is available for each feature, the least important features are pruned from current set of features. That procedure is recursively repeated on the pruned set until the desired number of features to select is eventually reached[10].

For a default model, the Recursive Feature Elimination (RFE) has helped us select the following features: *sale, gdp, loan_form_credit, marriage_married, marriage_single, repeatedly_0, repeatedly_1, repeatedly_2, repeatedly_3, product_I12, region_Korea, clear_year_2016, clear_year_2018, clear_year_2019, beginning_year_2015, beginning_year_2016, ending_year_2016, ending_year_2018, ending_year_2019, ending_year_2020.*

For prepayment model: *LTV, duration, gdp, loan_form_credit, marriage_married, marriage_single, product_I3, product_I6, product_PI12, product_PI18, product_PI24, product_PI6, clear_year_2017, clear_year_2018, clear_year_2019, beginning_year_2016, beginning_year_2018, end_year_2018, ending_year_2019, ending_year_2020* have been selected.

[9] Source :https://docs.scipy.org/doc/scipy-0.13.0/reference/generated/scipy.special.Logit.html
[10] Source :https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFE.html

We could conclude that the time dimension dummy variables (*beginning year, clear year, end year dummies)* representing different time states are important to explain both default and prepayment probability. Such a conclusion implies that repayment risk may depend on different macro economical states. In addition, a client's marital status and local GDP index also seem to be signal determinants both in default and prepayment risk models.

Furthermore, clients' choices of different loan products may explain their prepayment decision better than in default model. Variable *repeatedly dummies* imply a client's seniority and also explain default probability, while *LTV* (*loan amount/auto value*) can to some extent shed light on a client's potential prepayment intention.

Model 4. Logit with recursive feature elimination without time dimension dummies.

We have discussed above that the repayment risk of client depends significantly on time dimension dummies. The objective of this analysis is to extract important determinants of repayment risk in order to serve a risk early warning mechanism. Considering the difficulty to forecast effectively the following time state, in this model, we will only focus on characteristic variables (client, auto loan and automobile.) By excluding time dimension dummies from the basic Logit model, the RFE process has helped us select the following features:

For the default model: *LTV, duration, sale, loan_form_credit, gender_F, marriage_married, marriage_single, repeatedly_0, repeatedly_1, repeatedly_2, repeatedly_3, product_I12, product_I6, product_PI12, product_PI18, product_PI24, region_China, region_Korea, region_UK, nature_private*.

For the prepayment model: *duration, sale, gdp, loan_form_credit, gender_F, marriage_married, marriage_single, repeatedly_0, repeatedly_1, repeatedly_2, product_I12, product_I3, product_PI12, product_PI18, product_PI24, licence_wuxi, region_China, region_Korea, nature_private.*

After the exclusion of year's dummies, the importance of the variable region of automobile manufactures surfaced, which is in accordance with the study of Agarwal et al. (2008): loans on most Japanese automobiles have a lower probability of default. Clients owning automobiles with medium to low brand image (domestic or Korean) may show more repayment risk. Furthermore, loans with longer duration may face more cash flow pressure and thus explain prepayment and default events. Collateral cars as personal property have higher risk exposure than that as corporate property. In addition, this model differs from last with time dimension dummies, as clients' choices on different loan finance products reveal more importance in default model.

VI. **Empirical Results**

Table 8 presents coefficients and their t-statistics of the default and prepayment competing risks models using the 4 methods described above.

As outlined earlier, we first estimate the baseline Probit and Logit function then prune parameters in order to filter out those who have the most explanatory power. Competing risks models present the maximum likelihood estimates of the parameters for the previous section's specifications. We have kept the same regressors for the Probit and Logit and for the default and prepayment model in order to illustrate that some variables can be significant in one model but not in the other. As we will see, several significant variables in the default probability equation are not significant in the prepayment probability equation and vice versa. Also, several important variables in the default probability equation picked by recursive feature elimination process are not necessarily being picked in the prepayment probability equation and vice versa. Our following interpretation only focuses on significant variables.

Table 8. Competing risks models of default and prepayment

Variable	PROBIT		LOGIT		LOGIT REF		LOGIT REF without year's dummies	
	Default	**Prepay.**	**Default**	**Prepay.**	**Default**	**Prepay.**	**Default**	**Prepay.**
Const	-14.348	-9.948	-49.688	-34.129	-20.779	1.377	-18.893	1.683
	(-0.004)	(-0.003)	(0.000)	(0.000)	(-0.001)	(1.274)	(-0.004)	(1.597)
Borrower's characteristics variables								
Age	-0.001	0.000	-0.001	0.005	-	-	-	-
	(-0.403)	(0.089)	(-0.286)	(0.448)				
Gender _F	0.032	-0.209**	0.045	-0.139	-	-	-0.516***	-0.253
	(0.577)	(-2.070)	(0.472)	(-0.664)			(-5.624)	(-1.374)
Marriage _married	-0.430***	0.311*	-0.736***	0.706*	-0.672***	0.598*	-0.832***	0.742**
	(-4.629)	(1.710)	(-4.582)	(1.823)	(-4.579)	(1.697)	(-5.674)	(2.052)
Marriage _single	-0.227**	0.208	-0.375*	0.558	-0.546***	0.692*	-0.688***	0.749*
	(-2.022)	(0.948)	(-1.951)	(1.191)	(-3.178)	(1.697)	(-4.007)	(1.815)
Province _ native	0.184***	-0.090	0.285***	-0.213	-	-	-	-
	(2.982)	(-0.849)	(2.786)	(-0.963)				
Loan's characteristics variables								
Loan	-0.006***	0.003	-0.010***	0.006	-	-	-	-
	(-4.207)	(1.155)	(-4.235)	(1.051)				
LTV	0.876***	-0.999**	1.440***	-2.180**	-	-1.423***	0.655***	-0.600

Table 8. (continued)

	(3.604)	(-2.099)	(3.544)	(-2.250)		(-2.770)	(2.610)	(-1.176)
Duration	0.125	0.228	0.193	0.113	-	-1.064***	0.214***	-1.130***
	(1.451)	(1.428)	(1.336)	(0.343)		(-21.154)	(20.953)	(-22.791)
Monthly _pay	-0.003*	-0.005**	-0.005*	-0.006	-	-	-	-
	(-1.819)	(-2.203)	(-1.652)	(-1.209)				
Clear date	0.000	-0.024***	0.001	-0.045***	-	-	-	-
	(0.016)	(-4.520)	(0.157)	(-4.050)				
Beginning date	0.002	0.000	0.004	-0.011	-	-	-	-
	(0.761)	(0.023)	(0.690)	(-0.805)				
Ending date	-0.005***	0.024***	-0.009***	0.054***	-	-	-	-
	(-5.162)	(7.755)	(-5.287)	(7.353)				
Loan form _credit	-0.399***	0.558*	-0.747***	1.081*	-0.742***	1.547***	-0.803***	1.338***
	(-2.609)	(1.730)	(-2.884)	(1.704)	(-3.410)	(3.634)	(-3.613)	(3.055)
Repeatedly _0	7.927	0.389	27.756	2.142	22.194	-	19.181	-0.045
	(0.006)	(0.207)	(0.000)	(0.391)	(0.001)		(0.004)	(-0.072)
Repeatedly _1	7.818	0.407	27.589	2.218	22.009	-	19.032	-0.152
	(0.006)	(0.216)	(0.000)	(0.405)	(0.001)		(0.004)	(-0.240)
Repeatedly _2	8.300	0.217	28.416	1.516	22.633	-	19.524	-1.091
	(0.006)	(0.115)	(0.000)	(0.276)	(0.001)		(0.004)	(-1.459)
Repeatedly _3	8.201	0.462	28.214	2.206	22.714	-	20.072	-
	(0.006)	(0.237)	(0.000)	(0.395)	(0.001)		(0.004)	
Product _I12	-7.342	4.929	-17.849	22.392	-13.340	-	-21.945	7.282
	(0.000)	(0.001)	(-0.008)	(0.000)	(-0.061)		(-0.002)	(6.093)
Product _I3	-0.452	5.624	-0.858	22.837	-	-5.035***	-	-5.135
	(-1.338)	(0.001)	(-1.373)	(0.000)		(-5.799)		(-12.565)
Product _I6	0.165	6.826	0.135	25.321	-	0.973	-0.514	-
	(0.407)	(0.002)	(0.185)	(0.000)		(1.017)	(-1.632)	
Product _PI12	0.008	5.208	-0.101	22.413	-	6.974***	-1.313***	7.121***
	(0.018)	(0.001)	(-0.131)	(0.000)		(7.911)	(-13.464)	(17.768)
Product _PI18	0.768	3.181	1.229	18.602	-	12.125***	-1.773***	12.689***
	(1.337)	(0.001)	(1.247)	(0.000)		(11.612)	(-12.274)	(19.652)
Product _PI24	2.650***	-0.484	4.224***	11.323	-	15.282***	-1.906***	15.032***
	(3.366)	(0.000)	(3.158)	(0.000)		(9.582)	(-6.621)	(15.336)
Product _PI6	0.021	6.591	-0.077	25.358	-	0.394	-	-
	(0.058)	(0.002)	(-0.115)	(0.000)		(0.463)		
Automobile's characteristics variables								
Car value	0.005***	-0.003	0.008***	-0.006	-	-	-	-
	(4.354)	(-1.261)	(4.322)	(-1.246)				

Table 8. (continued)

Licence_wuxi	-0.087	-0.429	-0.169	-0.658	-	-	-	-0.307
	(-0.443)	(-1.293)	(-0.510)	(-1.027)				(-0.611)
Region_China	0.178**	-0.462***	0.316**	-0.732**	-	-	0.417***	-0.611***
	(2.261)	(-3.216)	(2.406)	(-2.446)			(4.245)	(-2.772)
Region_EU	0.085	-0.194	0.152	-0.267	-	-	-	-
	(1.240)	(-1.572)	(1.321)	(-1.063)				
Region_Japan	-0.082	-0.105	-0.122	-0.258	-	-	-	-
	(-1.210)	(-0.869)	(-1.085)	(-1.047)				
Region_Korea	0.525***	-0.282	0.929***	-0.498	0.592***	-	0.688***	-0.723*
	(4.952)	(-1.381)	(5.107)	(-1.143)	(3.920)		(4.429)	(-1.816)
Region_ UK	0.159	-0.207	0.253	-0.329	-	-	0.225	-
	(0.933)	(-0.650)	(0.870)	(-0.503)			(0.921)	
Nature_private	0.137	0.086	0.213	0.221	-	-	0.379***	0.465
	(1.539)	(0.523)	(1.446)	(0.637)			(2.765)	(1.561)
Economic conditions variables								
Income	-0.475***	0.651***	-0.693***	0.559*	-	-	-	-
	(-5.337)	(4.159)	(-4.516)	(1.835)				
Sale	7.458***	-13.060***	9.390***	-6.467	-15.495***	-	-1.842***	3.761
	(4.229)	(-4.227)	(2.986)	(-1.014)	(-8.971)		(-4.168)	(2.514)
GDP	-1.357***	4.839***	-1.616*	5.137***	4.997***	2.683***	-	0.522
	(-2.722)	(5.677)	(-1.843)	(2.947)	(7.448)	(7.530)		(0.502)
Time dimension dummies								
Clear year_2016	-0.726***	0.261	-1.283***	0.928*	-1.072***	-	-	-
	(-6.175)	(1.135)	(-6.485)	(1.853)	(-9.219)			
Clear year_2017	-0.004	-0.362	-0.063	-0.161	-	-1.120***	-	-
	(-0.024)	(-1.031)	(-0.201)	(-0.208)		(-4.802)		
Clear year_2018	0.919***	-0.529	1.515***	-0.373	2.473***	-1.707***	-	-
	(3.688)	(-1.191)	(3.588)	(-0.382)	(14.932)	(-3.541)		
Clear year_2019	1.221***	-0.261	1.953***	-0.286	3.744***	-2.066***	-	-
	(3.711)	(-0.475)	(3.536)	(-0.243)	(13.511)	(-3.041)		
Beginning year_2015	7.568	-0.118	24.111	0.353	0.601***		-	-
	(0.003)	(-0.204)	(0.001)	(0.276)	(4.056)			
Beginning year_2016	8.224	0.026	25.180	0.497	0.752***	0.833***	-	-
	(0.003)	(0.039)	(0.001)	(0.349)	(6.044)	(3.924)		
Beginning year_2017	8.399	0.163	25.454	0.334	-	-	-	-
	(0.003)	(0.220)	(0.001)	(0.209)				
Beginning year_2018	9.667	-0.084	27.539	-0.135	-	-0.508	-	-
	(0.003)	(-0.100)	(0.001)	(-0.075)		(-1.301)		

Table 8. (continued)

Beginning year_2019	10.547	-0.204	28.989	-1.047	-	-	-	-
	(0.004)	(-0.152)	(0.001)	(-0.241)				
Ending year_2016	0.991***	0.017	1.722***	-0.536	0.707***	-	-	-
	(6.887)	(0.052)	(7.142)	(-0.786)	(5.454)			
Ending year_2017	0.957***	0.487	1.603***	0.329	-	-	-	-
	(4.035)	(0.967)	(4.036)	(0.317)				
Ending year_2018	0.838***	0.302	1.329**	0.486	-1.323***	0.967**	-	-
	(2.712)	(0.484)	(2.546)	(0.370)	(-8.436)	(2.176)		
Ending year_2019	0.259	0.394	0.353	1.201	-2.893***	1.962***	-	-
	(0.661)	(0.525)	(0.536)	(0.755)	(-11.23)	(2.897)		
Ending year_2020	-1.187**	-0.384	-1.991**	-0.356	-4.463***	0.431	-	-
	(-2.077)	(-0.382)	(-2.005)	(-0.152)	(-9.980)	(0.261)		
Number of parameters	50	50	50	50	21	21	21	21
Number of observations	4722	3066	4722	3066	4722	3066	4738	3134
Pseudo R^2	0.1933	0.6904	0.1954	0.7450	0.120	0.688	0.118	0.689
Log-likelihood	-2640.4	-657.92	-2633.4	-541.93	-2881.3	-663.62	-2897.0	-675.35
LLR p-value	9.826e-233	0.000	1.206e-235	0.000	4.7214e-153	0.0000	3.9028e-151	0.0000

Note: The table provides the coefficient values and the t-statistics (below the coefficient value, in the parenthesis).
* significant at the 10 percent level, ** significant at the 5 percent level, *** significant at the 1 percent level.

Concerning the first two baseline models Probit and Logit that we represented above, as we could observe, the two models produce similar results, give almost the same significant variables, approximate coefficients and the same signs. Among all variables, *Gender_F* is significant at 5% significant level to reduce prepayment probability in Probit model while it is not significant in prepayment Logit model. *Clear Year_2016* is significant at 1% significant level to raise prepayment probability in Logit model while it is not significant in prepayment Probit model. Log-likelihood of default and prepayment models in Probit models are: default model 2640.4; prepayment model -657.92, and those in Logit model correspond to: default model 2633.4; prepayment model -541.93. Logit models have higher Log-likelihood overall which illustrates that it has more explanatory power than the Probit. Based on the log-likelihood statistic for these models, the pseudo R^2 [11]of default and prepayment models in Probit models are: default model 19.33 percent; prepayment model 69.04 percent, and those in Logit model correspond to: default model 19.54

[11] The pseudo R^2 is calculated from the ratio of the model log-likelihood statistic to the restricted model log-likelihood statistic, where the restricted model have only an intercept term as regressor.

percent; prepayment model 74.5 percent. All the LLR p-value are approximated to zero, which confirm the validity of the models. Pseudo R^2 of prepayment models are all higher than these of default models, which is understandable as default model has such imbalanced data.

For time dimensional dummies variables, most of *clear year* and *ending year dummies* are significant in Probit and Logit default model. Referencing with 2015 (reference group), if loans actually closed in 2016 (*clear year_2016*) their default probability is reduced, and if loans actually closed in 2018 (*clear year_2018*) their default probability is increased. This result confirms Li et al. (2018) that interest rates of the loan positively correlates with default probability. Figure 6 presents a candle chart of China 10-year Treasury. It can be observed that China 10-year Treasury rate, which is considered as risk-free rate, dropped in 2016 and increased heavily in 2018. However, in the prepayment model, dummy variables mentioned above are not all significant. The default model seems to depend more on different time state.

Figure 6. China 10-year Treasury - Candle chart

For borrower's characteristics variables, females have less intention to prepay. Referencing with divorced clients, married or single client have less default probability and higher prepayment intention. Native clients have higher default probability but not significant in prepayment model.

For a loan's characteristics variables, a loan with higher amount reduces default probability but is not significant in prepayment model. Higher *monthly_pay* will reduce default probability and prepayment probability at the same time, but it is weeded out in pruned models. For count variable *ending date*, longer *ending date* may imply the maturity of institution's lending business which reduces default probability and raises prepayment probability. In accordance with the study of Agarwal et al. (2005), the impact of the *LTV* (loan amount / car value) ratio also follows the anticipated pattern. The positive coefficient in the default

model indicates that the probability of default increases as the *LTV* increases. Since these loans are positive amortizing loans, an increase in the *LTV* implies that the underlying asset value has declined. On the prepayment side, the negative coefficient suggests the opposite effect. Also, as we have discussed before, compared with clients who have a car that is collateral immobilized in intuitional lender, *loan form_credit* has less probability to default and higher probability to prepay. Comparing with clients who choose equal total loan payments with 3 months maturity (*product_PI3*, the reference group), clients who choose the same equal total loan payment method but with longer maturity have a higher default probability and a lower prepayment probability.

For car characteristics variables, clients with a car having a higher value have higher probability to default, but this is not significant in repayment model. Furthermore, clients with a car manufactured in China have a higher default probability and a lower probability for prepayment. The same phenomenon can be observed in cars manufactured in Korea, where manufacturers may have relatively low-mid brand image.

Finally, we note that the local *GDP* and *income* (the per capita disposable income of provincial residents) are significantly negative in the default model and positive in the prepayment model. In fact, they are both proxies for local economic conditions, with higher rates implying better economic conditions. Thus, the negative coefficient in the default model implies that during periods of greater economic uncertainty, the probability of auto loan default increases; and the positive coefficient in the prepayment model implies that during periods of economic recession, the probability of prepayment decreases. However, we note that *sale* (retail sales of consumer goods) in the Probit and Logit non-pruned models have inverse signs compared with *GDP* and *income*. It may be because *sales* have strong correlation with *GDP* (0.973) and *income* (0.966). This problem is corrected in pruned model, where *sales* has logical sign.

1. Predictive performance of default and prepayment models on training dataset

Table 9 provides confusion matrixes representing Probit and Logit models' predictive performance on training dataset.

In the matrixes, FP is a false positive corresponding to a type I error, and FN is a false negative corresponding to a type II error. Accuracy criteria is calculated by: (TN+TP) / (TN+FP+FN+TP). The default model in Probit or Logit achieves accuracy 72% and 73% respectively, while the prepayment model achieves accuracy 95% and 96% respectively. We conclude that the Logit model is more accurate than the Probit model in predictive performance.

Table 9. Models' predictive performance on training dataset

Default Probit				Default Logit			
		Predicted				Predicted	
		negative	positive			negative	positive
Actual	negative	TN: 1757	FP: 604	Actual	negative	TN: 1783	FP: 578
	positive	FN: 704	TP: 1657		positive	FN: 702	TP: 1659
Accuracy	**0.72**			**Accuracy**	**0.73**		
Prepayment Probit				**Prepayment Logit**			
		Predicted				Predicted	
		negative	positive			negative	positive
Actual	negative	TN: 1430	FP: 604	Actual	negative	TN: 1474	FP: 59
	positive	FN: 49	TP: 1484		positive	FP: 59	FP: 59
Accuracy	**0.95**			**Accuracy**	**0.96**		

2. Marginal Effects

As we have discussed above, since the dependent variable is a nonlinear function of the regressors, the coefficient on X has no simple interpretation. In Probit regression, one-unit change in x_i leads to a β_i change in the z-score of *Y*. In our case, for example, a unit of one thousand CNY increase in loan amount will reduce the z-score of Pr $(default = 1)$ by 0.006.

Marginal effects describe an additive change in Pr $(Y = 1)$ for change in X_i holding other variables at specific values. In nonlinear function, this expression depends on not just β_i, but on the value of x_i and all other variables in the regression equation. Here we compute AME: the average of the marginal effects at each observation "overall."

Table 10. presents marginal effects of default and prepayment models. Using the case of explanatory variable *loan amount*, this marginal effect could be explained by holding other regressors in specific value, an additional unit of one thousand CNY increase in loan amount reduces the probability of default by 0.19 percent points.

For dummy variables, the marginal effect is the change in probabilities treating binary variables as changing from 0 to 1. With the example of *marriage _married* in the default model, if client is married, his or her default probability will reduce by 0.14 percent points, holding other regressors in specific value.

For count variable, *ending date* for example, the marginal effect is the change in probabilities when each observation is increased by one. In the default Probit model, this means that one additional day in *ending date* reduces the probability of default by 0.17 percent points, holding other regressors in specific value.

We could also notice the important marginal effects of loan products chosen by clients. In the Logit REF model, holding other regressors in specific value, referencing with clients who choose equal total loan payments with 3 months' maturity (*product_PI3*, the reference group), clients who choose the same equal total loan payment method but with 18 months' maturity (*product_PI18)* raise the probability of default by 0.7844. Clients who choose the same repayment method but with 24 months' maturity (*product_PI24*) raise the probability of default by 0.5177.

Furthermore , the region of auto manufacturer also has significant marginal effects. In Logit REF without year's dummies model, referencing with region USA (the reference group), clients with a car manufactured in Korea (*Region_Korea*) raise default probability by 14 percentage points and reduce prepayment probability by 4 percentage points.

In addition to this, time dimension dummies, client *gender*, *LTV* (*loan-to-value*), *loan duration*, loan *form_credit*, *sale*, *gdp* are also surfaced, comparable to important margin effects on default and prepayment probability.

Table 10. Marginal effects of competing risks models

Variable	PROBIT		LOGIT		REF LOGIT		REF LOGIT without time year's dummies	
	Default	Prepay.	Default	Prepay.	Default	Prepay.	Default	Prepay.
Borrower's characteristics variables								
Age	-0.000	0.001	-0.000	0.000	-	-	-	-
	(-0.403)	(0.089)	(-0.286)	(0.448)				
Gender_F	0.010	0.013**	0.009	-0.007	-	-	0.110***	-0.016
	(0.577)	(-2.080)	(0.472)	(-0.664)			(-5.694)	(-1.376)
Marriage_married	-0.137***	0.024*	-0.140***	0.037*	-0.143***	0.039*	0.177***	0.048**
	(-4.658)	(1.713)	(-4.619)	(1.827)	(-4.617)	(1.700)	(-5.748)	(2.057)
Marriage_single	-0.072**	0.029	-0.071*	0.029	-0.116***	0.045*	0.146***	0.048*
	(-2.025)	(0.948)	(-1.954)	(1.192)	(-3.191)	(1.701)	(-4.032)	(1.818)
Province_native	0.059***	0.014	0.054***	-0.011	-	-	-	-
	(2.991)	(-0.85)	(2.795)	(-0.964)				
Loan's characteristics variables								

Table 10. (continued)

Loan	-0.002***	0.000	-0.002***	0.000	-	-	-	-
	(-4.230)	(1.156)	(-4.264)	(1.051)				
LTV	0.279***	0.063**	0.274***	-0.114**	-	0.092***	0.139***	-0.039
	(3.619)	(-2.105)	(3.561)	(-2.256)		(-2.786)	(2.617)	(-1.177)
Duration	0.040	0.021	0.037	0.006	-	0.069***	0.046***	0.073***
	(1.453)	(1.432)	(1.336)	(0.343)		(-37.143)	(25.693)	(-49.891)
Monthly _pay	-0.001*	0.000**	-0.001*	-0.000	-	-	-	-
	(-1.821)	(-2.213)	(-1.654)	(-1.211)				
Clear date	0.000	0.001***	0.000	-0.002***	-	-	-	-
	(0.016)	(-4.592)	(0.157)	(-4.105)				
Beginning date	0.001	0.001	0.001	-0.001	-	-	-	-
	(0.762)	(0.023)	(0.690)	(-0.806)				
Ending date	-0.002***	0.000***	-0.002***	0.003***	-	-	-	-
	(-5.203)	(7.942)	(-5.339)	(7.695)				
Loan form _credit	-0.127***	0.043*	-0.142***	0.057*	-0.158***	0.100***	0.171***	0.086***
	(-2.613)	(1.734)	(-2.893)	(1.708)	(-3.425)	(3.694)	(-3.629)	(3.093)
Repeatedly _0	2.528	0.251	5.277	0.112	4.717	-	4.076	-0.003
	(0.006)	(0.207)	(0.000)	(0.391)	(0.001)		(0.004)	(-0.072)
Repeatedly _1	2.493	0.251	5.245	0.116	4.678	-	4.044	-0.010
	(0.006)	(0.216)	(0.000)	(0.405)	(0.001)		(0.004)	(-0.240)
Repeatedly _2	2.647	0.252	5.402	0.079	4.811	-	4.149	-0.070
	(0.006)	(0.115)	(0.000)	(0.276)	(0.001)		(0.004)	(-1.461)
Repeatedly _3	2.615	0.260	5.364	0.115	4.828	-	4.265	-
	(0.006)	(0.237)	(0.000)	(0.395)	(0.001)		(0.004)	
Product _I12	-0.500	515.426	-0.501	1.170	-0.501	-	-0.502	0.470
	(-55.422)	(0.001)	(-78.648)	(0.000)	(-74.268)		(-74.754)	(6.270)
Product _I3	-0.144	515.426	-0.163	1.193	-	0.326***	-	-0.331
	(-1.339)	(0.001)	(-1.374)	(0.000)		(-5.886)		(-13.518)
Product _I6	0.053	515.426	0.026	1.323	-	0.063	-0.109	-
	(0.407)	(0.002)	(0.185)	(0.000)		(1.018)	(-1.634)	
Product _PI12	0.003	515.426	-0.019	1.171	-	0.451***	0.279***	0.459***
	(0.018)	(0.001)	(-0.131)	(0.000)		(8.290)	(-14.478)	(24.450)
Product _PI18	0.245	515.426	0.234	0.972	-	0.784***	0.377***	0.818***
	(1.337)	(0.001)	(1.248)	(0.000)		(13.006)	(-13.032)	(30.984)
Product _PI24	0.845***	491.501	0.803***	0.503	-	0.518***	0.405***	0.970***
	(3.379)	(0.000)	(3.169)	(0.001)		(89.985)	(-6.729)	(19.355)
Product _PI6	0.007	515.426	-0.015	1.325	-	0.026	-	-
	(0.058)	(0.002)	(-0.115)	(0.000)		(0.463)		
Automobile's characteristics variables								

Table 10. (continued)

Car value	0.002***	0.000	0.002***	-0.000	-	-	-	-
	(4.379)	(-1.262)	(4.353)	(-1.246)				
Licence_wuxi	-0.028	0.044	-0.032	-0.034	-	-	-	-0.0198
	(-0.443)	(-1.294)	(-0.510)	(-1.028)				(-0.611)
Region_China	0.057**	0.019***	0.060**	-0.038**	-	-	0.089***	-0.039***
	(2.264)	(-3.243)	(2.411)	(-2.452)			(4.274)	(-2.785)
Region_EU	0.027	0.016	0.029	-0.014	-	-	-	-
	(1.240)	(-1.576)	(1.322)	(-1.064)				
Region_Japan	-0.026	0.016	-0.0232	-0.014	-	-	-	-
	(-1.210)	(-0.869)	(-1.085)	(-1.048)				
Region_Korea	0.168***	0.027	0.177***	-0.026	0.126***	-	0.146***	-0.047*
	(4.985)	(-1.383)	(5.156)	(-1.143)	(3.944)		(4.462)	(-1.819)
Region_UK	0.051	0.042	0.048	-0.017	-	-	0.048	-
	(0.933)	(-0.65)	(0.870)	(-0.503)			(0.922)	
Nature_private	0.044	0.022	0.041	0.012	-	-	0.081***	0.030
	(1.541)	(0.523)	(1.447)	(0.638)			(2.774)	(1.563)
Economic conditions variables								
Income	-0.151***	0.020***	-0.132***	0.029*	-	-	-	-
	(-5.394)	(4.234)	(-4.552)	(1.838)				
Sale	2.378***	0.403***	1.785***	-0.338	-3.294***	-	0.391***	0.243
	(4.262)	(-4.322)	(2.996)	(-1.015)	(-9.256)		(-4.191)	(1.497)
GDP	-0.433***	0.110***	-0.307*	0.268***	1.062***	0.174***	-	0.034
	(-2.731)	(5.861)	(-1.846)	(2.954)	(7.613)	(7.675)		(0.502)
Time dimension dummies								
Clear year_2016	-0.231***	0.031	-0.244***	0.049*	-0.228***	-	-	-
	(-6.24)	(1.136)	(-6.586)	(1.856)	(-9.555)			
Clear year_2017	-0.001	0.047	-0.012	-0.008	-	0.073***	-	-
	(-0.024)	(-1.032)	(-0.201)	(-0.208)		(-4.866)		
Clear year_2018	0.293***	0.059	0.288***	-0.020	0.526***	0.110***	-	-
	(3.704)	(-1.193)	(3.607)	(-0.382)	(16.298)	(-3.565)		
Clear year_2019	0.389***	0.073	0.371***	-0.015***	0.796***	-0.134***	-	-
	(3.729)	(-0.475)	(3.553)	(-0.243)	(14.515)	(-3.060)		
Beginning year_2015	2.413	0.077	4.584	0.018	0.128***	-	-	-
	(0.003)	(-0.204)	(0.001)	(0.276)	(4.082)			
Beginning year_2016	2.622	0.088	4.787	0.026	0.160***	0.054***	-	-
	(0.003)	(0.039)	(0.001)	(0.349)	(6.129)	(3.965)		
Beginning year_2017	2.678	0.099	4.839	0.017	-	-	-	-
	(0.003)	(0.220)	(0.001)	(0.209)				
Beginning year_2018	3.082	0.112	5.235	-0.007	-	-0.033	-	-
	(0.003)	(-0.100)	(0.001)	(-0.075)		(-1.302)		

Table 10. (continued)

Beginning year_2019	3.363	0.176	5.511	-0.054	-	-	-	-
	(0.004)	(-0.153)	(0.001)	(-0.243)				
Ending year_2016	0.316***	0.045	0.327***	-0.028	0.150***	-	-	-
	(6.981)	(0.052)	(7.277)	(-0.786)	(5.521)			
Ending year_2017	0.305***	0.067	0.305***	0.017	-	-	-	-
	(4.055)	(0.968)	(4.059)	(0.317)				
Ending year_2018	0.267***	0.083	0.253**	0.026	-0.281***	0.063**	-	-
	(2.718)	(0.484)	(2.553)	(0.370)	(-8.641)	(2.182)		
Ending year_2019	0.082	0.100	0.067	0.063	-0.615***	0.127***	-	-
	(0.661)	(0.525)	(0.536)	(0.756)	(-11.775)	(2.915)		
Ending year_2020	-0.379**	0.134	-0.379**	-0.019	-0.949***	0.0279	-	-
	(-2.079)	(-0.382)	(-2.008)	(-0.152)	(-10.348)	(0.261)		

Note: The table provides the average marginal effects: dy/dx (at the overall) and the t-statistics (below the coefficient value, in the parenthesis). * significant at the 10 percent level, ** significant at the 5 percent level, *** significant at the 1 percent level.

3. **Predictive performance of default and prepayment models on test dataset**

We examine here the models' predictive performance on test dataset by computing *precision*, *recall*, *F-measure* and *support:*

- The precision is the ratio TP / (TP + FP) where TP is the number of true positives and FP the number of false positives. The precision is intuitively the ability of the classifier not to label as positive a sample that is negative.
- The recall is the ratio TP / (TP + FN) where TP is the number of true positives and FN the number of false negatives. The recall is intuitively the ability of the classifier to find all the positive samples.
- The F-beta score can be interpreted as a weighted harmonic mean of the precision and recall, where a F-beta score reaches its best value at 1 and worst score at 0.
- The F-beta score weights recall more than precision by a factor of beta. beta = 1.0 means recall and precision are equally important.
- The support is the number of occurrences of each class in y_test.[13]

Table 11 provides a summary of the criteria described above. Both the default and prepayment model have an accuracy above 90%, which gives us strong predictive power of these two models. The default model predict non-defaulters more precisely than defaulters, which is not surprising as we have such an

[13] Source : https://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision_recall_fscore_support.html

imbalanced dataset. The prepayment model has generally a good prediction with slight lower recall ratio to find clients who prepay than those who do not.

Table 11. Models' predictive performance on test dataset

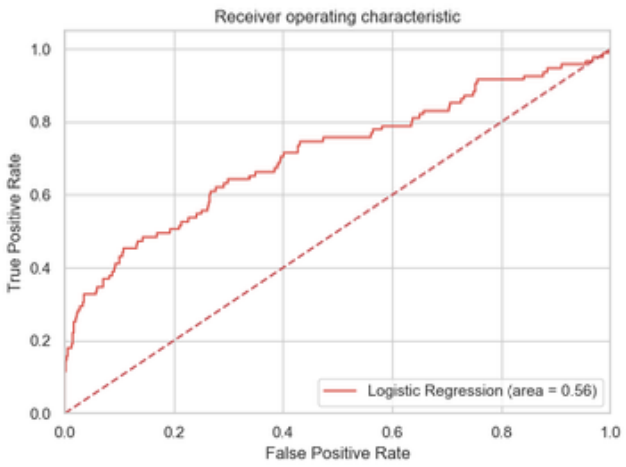
Default model				Prepayment model			
		Predicted				Predicted	
		negative	positive			negative	positive
Actual	negative	TN: 1004	FP: 0	Actual	negative	TN: 652	FP: 22
	positive	FN: 84	TP: 11		positive	FN: 36	TP: 389
Accuracy	0.92			**Accuracy**	0.95		

Default model					Prepayment model				
	Precision	**Recall**	**F1-Score**	**Support**		**Precision**	**Recall**	**F1-score**	**Support**
0	0.92	1	0.96	1004	0	0.95	0.97	0.96	674
1	1	0.12	0.21	95	1	0.95	0.92	0.93	425
Accuracy			0.92	1099	**Accuracy**			0.95	1099
Macro avg	0.96	0.56	0.58	1099	**Macro avg**	0.95	0.94	0.94	1099
Weighted avg	0.93	0.92	0.89	1099	**Weighted avg**	0.95	0.95	0.95	1099

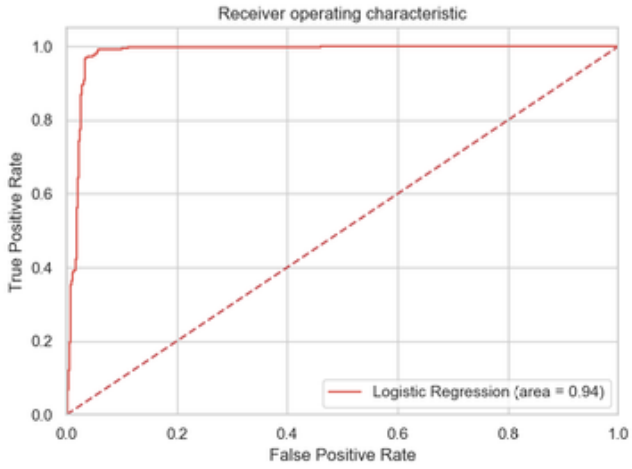
The weighted average in the table above can be interpreted as: Of the entire default test set, 93% of predicted defaulters were actual defaulters. Of the entire default test set, 92% of the actual defaulters were well predicted. Similarly, of the entire prepayment test set, 95% of predicted prepaid clients were actual prepaid. Of the entire prepayment test set, 95% of the actual prepaid clients are well predicted.

Figure 7. ROC accuracy score curve for logistic models of default and prepayment

ROC for Default model



ROC for prepayment model



ROC Curves summarize the trade-off between the true positive rate and false positive rate for a predictive model using different probability thresholds. It is a plot of the false positive rate (x-axis) versus the true positive rate (y-axis) for a number of different candidate threshold values between 0.0 and 1.0. To put it another way, it plots the false alarm rate versus the hit rate as it summarizes how often a positive class is predicted when the actual outcome is negative[14]. We could summarize the model skill by the area under the curve (AUC). A greater area under the curve indicates a greater model prediction accuracy skill. Compared directly in general or for different thresholds, for default logistic model, a threshold of 0.1389 provides the best trade-off between true positive rate and false positive rate. For prepayment logistic model, a threshold of 0.2595 provides the best result. The two optimal thresholds are found at the cut-off point that would be where the true positive rate is high and the false positive rate is low. The area under the prepayment logistic model is much bigger than the default logistic model, which means the prepayment logistic model has a relatively stronger predictive skill.

[14] Source: https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-classification-in-Python/

VII. Conclusion

In this study we use a unique data set from an automobile lending business company to isolate the determinant variables that affect default and prepayment risk probability in automobile loan market. The main objective is to improve the company's automobile lending business's maximization by building an ex-ante risk control system before the loan's releasing to predict accurately the repayment events and to reduce management cost.

The data set served in this study consisting of 9 variables and 4557 auto loan records of company between October 2014 to August 2019. After the needed data cleaning process and variable transformation, 50 explanatory variables have been recognized. These variables could be regrouped in 5 groups: borrower's characteristics, auto loan's characteristics, automobile's characteristics, local economic condition, and time dimension dummies variables. We adopt the competing risk framework to estimate jointly the default probability and prepayment probability in two baseline Probit and Logit models, then apply recursive feature elimination process to filter out the most powerful explanatory variables. Using these variables, the predictive power in the test data set achieved 92% accuracy in the default model and 95% accuracy in the prepayment model.

Our empirical results show that the variables that affirmed in literature to predict default and prepayment continue to perform as expected. Where we could note: females borrowers have less intention to prepay. Referencing with divorced clients, married or single clients have less default probability and higher prepayment intention. Longer loan duration increases default probability and decrease prepayment probability. *LTV* (loan amount/car value) rate positively correlate with default probability while negatively correlated with prepayment probability. Local economic conditions (*GDP, income*) are significantly negative correlated with default probability while positive correlated with prepayment probability.

Our results also provide evidence that the Logit model and the Probit model generate similar significant variables and approximate coefficients. In our case, the Logit model has a relatively better predictive power. Significant variables that affect default probability and prepayment probability are not the same. For example, a native client has a higher default probability, but this is not significant in the prepayment model. Besides, the most important variables selected in the default probability are not necessarily selected in the prepayment probability equation and vice versa. Therefore, limiting the estimation of default probability to calculate repayment risk cost is not enough to calculate accurately repayment risk event cost. We also observe that default and prepayment probability significatively depend on different time states; in addition, it could be observed and confirmed that interest rates positively correlated with default probability.

Other self-selection evidence is also provided: the choice of loan products with different repayment methods and maturity reveals information about a client's propensity to default or prepay. Clients owning automobiles with medium to low-mid brand image (domestic or Korean) have a higher default probability and a lower probability for prepayment, and if we regard the origin and value of an automobile as a proxy of a client's wealth, this may refer to a wealth or income effect on default and prepayment probability. Furthermore, clients who have an auto loan supported by credit have a lower default probability and a higher prepayment probability.

Our data limited the analysis in this study. As the lending company only recorded accepted clients, clients who constitute a high risk are refused without any information recorded. For further extension, we could introduce information for those who have not received a loan and consider their past experience in the loan market if data are available. Furthermore, our conclusion might be only company specific, in that it may permit to reduce the heterogeneity problem once we can access the same type of data from auto loan industry.

Another extension is that we could consider default and prepayment decision as options held by clients. Let's take the default option as an example: if a client chooses a float interest rate for loan repayment, the market value of the automobile stands as the strike price and the total repayment amount stands as underlying assets. Once the total repayment amount of the loan rises with the increasing interest rate and exceeds the market value of the automobile, the client may make a default decision. Within this framework, the value of a car depreciates constantly, thus implying a float strike price. We have to apply numerical methods to price the default option, such as the finite difference method and the Monte Carlo simulation, which could serve as a better cost calculation and loan pricing from the perspective of borrowers. However, as the data set in this study is cross-sectional, loan's market value, collateral asset's market value, automobile's depreciation rate and cost information are not available in the time line. Consequently, we could not achieve a deeper analysis. Modeling the termination probability of auto loans in option's framework may serve to create a more accurate method of loan pricing and repayment event's cost calculation.

The last extension would be to introduce information on clients who have both revolving line of credit and an auto loan. According to D'astous, Dionne and Bergerès (2015), there exists strong evidence of dependence between the two financial instruments. Credit line utilization can reduce the default probability of loans, while a loan default can increase the utilization of credit lines. This leads to further studies on the joint modeling of credit line utilisation and default probability by using instrumental variables in order to deal with endogeneity in the system of joint model equation. Following D'astous, Dionne and Bergerès (2015), to instrument the default probability on auto loans, we could use the remaining loan term and

existing collateral in auto loans. Considering the fact that the existence of loan collateral does not affect the borrower's liquidity needs, and the remaining term of the loan is externally dependent on the time since the term loan was originally signed, these instruments should be effective. With the same argument, to instrument credit line utilization, the number of months since the credit line account was opened, and the dummy variable that indicates the existence of collateral on the credit line will be effective instrumental variables. Given the fact that the two financial products of credit line and auto loan are treated independently by different financial institutions, once credit line data is available, the correlations of credit line utilization and default probability should allow lenders to more effectively apply credit risk diversification. Financial institutions can manage the risks of borrowers by considering the significant dependencies between their various financial instruments, create a more diversified loan portfolio, and in turn reduce the minimum capital reserve required by regulators.

VIII. References

Agarwal, Sumit and Ambrose, Brent W. and Chomsisengphet, Souphala. (2008) "Determinants of automobile loan default and prepayment" Economic Perspectives, Vol. 32, No. 3, 2008

Agarwal, Sumit and Ambrose, Brent W. and Chomsisengphet, Souphala. (2005) "Asymmetric Information and the Automobile Loan Market". Available at SSRN: https://ssrn.com/abstract=754828.

Altman, E.I., R.B. Avery, R.A. Eisenbeis and J.F. Sinkey. (1981) Application of classification techniques in business, banking and finance (JAI Press, Greenwich, CT).

Ambrose, Brent & Sanders, Anthony. (2003) "Commercial Mortgage-Backed Securities: Prepayment and Default." The Journal of Real Estate Finance and Economics. 26. 179-96. 10.2139/ssrn.299298

Bergerès, Anne-Sophie & d'Astous, Philippe & Dionne, Georges. (2015). "Is there any dependence between consumer credit line utilization and default probability on a term loan? Evidence from bank-customer data", Journal of Empirical Finance, Elsevier, vol. 33(C), pages 276-286.

Boyes, W.J., D.L. Hoffman and S.A. Low. (1989) An econometric analysis of the bank credit scoring

Dionne, Georges. (2019) Corporate Risk Management: Theories and Applications Wiley; 1st edition. ASIN: B07R6B5V8B.

Dionne, Georges & Artis, Manuel & Guillen, Montserrat. (1996) "Count data models for a credit scoring system," Journal of Empirical Finance, Elsevier, vol. 3(3), pages 303-325, September.

Dionne, Georges and C. Vanasse. (1993) "Automobile insurance ratemaking in the presence of asymmetrical information." Journal of Applied Econometrics 7, 149-165.

Dionne, Georges & Liu, Ying. (2017) "Effects of Insurance Incentives on Road Safety: Evidence from a Natural Experiment in China," Working Papers 17-1, HEC Montreal, Canada Research Chair in Risk Management, revised 15 Oct 2019.

Einav, Liran & Jenkins, Mark & Levin, Jonathan. (2013) "The Impact of Credit Scoring on Consumer Lending." The RAND Journal of Economics. 44. 10.1111/1756-2171.12019.

Guyon, I., J. Weston, S. Barnhill, and V. Vapnik. (2002) "Gene selection for cancer classification using support vector machines." Machine Learning 46 (1): 389–422.

Gross, D.B., and N.S. Souleles. (2002) "An Empirical Analysis of Personal Bankruptcy and Delinquency." Review of Financial Studies 15(1): 319–47.

Heitfield, Erik and Tarun Sabarwal. (2004) “What Drives Default and Prepayment on Subprime Auto Loans?” The Journal of Real Estate Finance and Economics, 29(4), 457-477. http://dx.doi.org/10.1023/B:REAL.0000044023.02636.e6.

Jaffee, D.M. and T. Russell. (1976) “Imperfect information, uncertainty and credit rationing.” Quarterly Journal of Economics, 651-666.

Liu, Huan & Ma, Lin & Zhao, xi & Zou, Jianhua. (2018) An Effective Model Between Mobile Phone Usage and P2P Default Behavior. 10.1007/978-3-319-93701-4_36.

Lingnan, Lin. (2019) Gender effect on the default risk in peer-to-peer lending markets: The case of the largest Chinese platform. Risk Governance and Control: Financial Markets & Institutions. 9. 8-22. 10.22495/rgcv9i3p1.

Li, Zhiyong & Li, Ke & Yao, Xiao & Wen, Qing. (2018) “Predicting Prepayment and Default Risks of Unsecured Consumer Loans in Online Lending.” Emerging Markets Finance and Trade. 55. 10.1080/1540496X.2018.1479251.

Mealli, F., and S. Pudney. (1996) “Occupational Pensions and Job Mobility in Britain: Estimation of a Random-Effects Competing Risks Model.” Journal of Applied Econometrics 11: 293–320.

Shumway, T. 2001. “Forecasting Bankruptcy More Accurately: A Simple Hazard Model.” Journal of Business 101–24.

Polena, Michal & Regner, Tobias. (2018) Determinants of Borrowers’ Default in P2P Lending under Consideration of the Loan Risk Class. Games. 9. 82. 10.3390/g9040082.

Steenackers, A. and M.J. Goovaerts. (1989) “A credit scoring model for personal loans.” Insurance: Mathematics and Economics 8, 31-34.

Stiglitz, J.E. and A. Weiss. (1981) “Credit rationing in markets with imperfect information.” American Economic Review 71, 393-410.

Suzuki, Yukiya. (2018) Credit Risk Analysis of Auto Loan in Latin America. 10.1007/978-3-319-92046-7_52.

Train, K., and C. Winston. (2004) “Vehicle Choice Behavior and the Declining Market Share of U.S. Automakers.” Working Paper, University of California, Berkeley.

Xuchen Lin, Xiaolong Li & Zhong Zheng. (2017) Evaluating borrower’s default risk in peer-to-peer lending: evidence from a lending platform in China, Applied Economics, 49:35, 3538-3545, DOI:10.1080/00036846.2016.1262526

Zhou, Guangyou & Zhang, Yijia & Luo, Sumei. (2018) P2P Network Lending, Loss Given Default and Credit Risks. Sustainability. 10. 1010. 10.3390/su10041010.