Risk and return prediction for pricing portfolios of non-performing consumer credit

Siyi Wang∗
School of Data Science
City University of Hong Kong
swang348-c@my.cityu.edu.hk

Xing Yan∗
Institute of Statistics and Big Data
Renmin University of China
xingshan@ruc.edu.cn

Bangqi Zheng
JD Digits
zhengbangqi@jd.com

Hu Wang
JD Digits
wanghu5@jd.com

Wangli Xu
Center for Applied Statistics and School of Statistics
Renmin University of China
wlxu@ruc.edu.cn

Qi Wu†
School of Data Science
City University of Hong Kong
qiwu55@cityu.edu.hk

ABSTRACT
We design a system for risk-analyzing and pricing portfolios of non-performing consumer credit loans. The rapid development of credit lending business for consumers heightens the need for trading portfolios formed by overdue loans as a manner of risk transferring. However, the problem is nontrivial technically and related research is absent. We tackle the challenge by building a bottom-up architecture, in which we model the distribution of every single loan’s repayment rate, followed by modeling the distribution of the portfolio’s overall repayment rate. To address the technical issues encountered, we adopt the approaches of simultaneous quantile regression, R-copula, and Gaussian one-factor copula model. To our best knowledge, this is the first study that successfully adopts a bottom-up system for analyzing credit portfolio risks of consumer loans. We conduct experiments on a vast amount of data and prove that our methodology can be applied successfully in real business tasks.

KEYWORDS
Consumer lending, credit portfolio risk, overdue loans, bottom-up system, dependence structure

∗Equal contribution.
†Corresponding author.

1 INTRODUCTION
In some countries, the recent rapid development of consumer credit lending business which lies in the center of FinTech industry [8], has changed a large population’s financial lifestyle and consumption patterns. It enables credit lending to those non-standard customers (to traditional commercial banks), and more flexible credit services are provided than in traditional credit card business. The credit providers are usually some big online platforms that have advantages in technology and data. They adopt many data-driven AI techniques to manage the whole workflow of the business.

While it has the potential of benefiting consumption promotion and activating small and medium-sized enterprises’ economic activities [16], it faces many challenges when running this business, among which the inherent risk is the core [35]. The most important challenge and technical problem in this business is the management of risks throughout the entire life stages of all loans, using the accumulated large amounts of data and the most advanced AI techniques. In fact, there have been lots of industry practices such as credit scorecards for managing the default-related risks, but some issues still remain unsolved and need the design of new tools.

While most existing research works and industry practices concentrate on the default risk of every single loan, an unresolved but extremely important problem is the risk of a large pool of loans, or a loan portfolio. Analyzing such type of risk is crucial and necessary in both the pricing and the regular risk management of non-performing loan portfolios, i.e., the loans that have been overdue. Such portfolios can form asset-backed securities, the trading of which can be used as a manner of risk transferring. This is in
great need when many lending providers have accumulated large amounts of risky loans and want to trade them for various purposes. Overdue loans are sensitive to economic status, or they share some common underlying economic risk factor. Only modeling well the risk of every single loan is far away from determining the overall risk of a loan portfolio. Thus, developing a model system that can incorporate both the modeling of single risk and the capturing of overall risk is essential for non-performing loan portfolios. Mathematically, this needs us to determine the marginal distributions of the single risk-related quantities as well as the dependence structure among them, which is the goal of this paper.

In this paper, we consider modeling both the repayment rate of each overdue loan in a portfolio, and the overall repayment rate of the whole portfolio, in the distribution sense. When a loan is not overdue, generally one will model its default probability. Otherwise, the overdue loan is treated as a defaulted asset and its repayment rate (or recovery rate, a term in the literature of corporate bonds) is considered to be modeled by us, which is a mixed variable of discrete and continuous. Moreover, in finance, the term risk refers to the uncertainty of a quantity related to one’s profit and loss, and is measured by some properties of the quantity’s distribution. Hence for the need of risk pricing, we will focus on the prediction of a variable’s distribution, which is very different from the single-value prediction in traditional machine learning research.

1.1 Literature Review

There has been a large body of literature focusing on the problem of credit scoring of the single customer or loan. Recent papers include [18], [11], [3], [2], [20], [26], etc., or the comprehensive review [38]. Techniques for credit scoring include logistic regression [19], support vector machines [5], [15], artificial neural networks [31], [28], [1], [2], and so on. Given the existing studies today, few of them take care of the non-performing credit, i.e., the overdue loans.

Moreover, from another point, very few papers considered the default relationships among a group of loans. What we can find is in [33], authors analyzed the retail portfolio credit risk with lagged macroeconomic risk drivers. [27] used simulation method to study the default conditions of consumer loans and concluded that consumer loans have a high covariance beta. [6] studied default contagious chains in the guarantee network in order to identify potential systemic financial risk.

In fact, studies on credit portfolio risk have existed for a long time, but the focus is on corporate bonds and the pricing of CDO (collateralized debt obligations). The examples include [25], [24], [39], [17], [32], [14], and the book [30]. The key technical problem of these works is similar to one of ours in this paper: capturing the dependence structure or correlations among loans. They all adopted a factor copula model to capture the correlation structure among the bonds in a basket. But different from the situation we face, traders in financial markets can imply the parameters of the model from the market CDO quotes using the pricing formula, while we must infer the model parameters from the data collected because there are no active trades happened or there is no such market. Given the data of overdue loans with unique characteristics such as extremely heterogeneous and ultra high-dimensional, the technical problem we need to solve is a challenging task with no attempts before as we know.

1.2 Contributions

To summarize, the contributions of our paper are three-folds:

1) We build a bottom-up architecture for estimating and predicting non-performing consumer credit portfolios’ repayment risk with overdue loan data. The data is of mixed type, extremely heterogeneous, and ultra high-dimensional. We tackle the challenge that few researchers have attempted before in the existing literature.

2) Specifically, we predict the marginal distribution of each single loan’s repayment with simultaneous quantile regression, deal with the mixed repayment rate issue with R-copula, model the correlation structure of loans via one-factor Gaussian copula, and finally use Monte Carlo simulation to obtain the distribution of the portfolio’s overall repayment rate.

3) We show in the experiments that our system can be applied successfully in real business tasks, including the pricing and the risk management of overdue loan portfolios. It may meet the needs of rapidly growing business of risk transferring.

2 PRELIMINARIES

2.1 Quantile Regression

Quantile regression [23] [29] is a powerful tool that enables us to look at the slices of the conditional distribution \( p(y|x) \) without any assumptions on the global distribution. Recently there have been very successful adoptions of multi-level quantile regression or its extension to financial forecasting problems, such as tail risk forecasting in financial markets [40] and loss prediction of defaulted bank loans [21]. We utilize multi-level quantile regression to approximate the full marginal distribution of each loan’s repayment rate, as shown in later sections.

Suppose a random variable \( Y \) has cumulative distribution function \( F_Y \), the \( \tau \)-quantile of \( Y \) is defined as \( Q_Y(\tau) = F_Y^{-1}(\tau), \tau \in (0, 1) \). If we further assume that for a fixed \( \tau \), the conditional quantile of \( Y \) given another variable \( X \) is a parameterized function: \( Q_Y(\tau|X=x) = f_\theta(x) \), then the function \( f_\theta \) can be estimated using quantile regression by minimizing the following expected loss:

\[
\min_\theta E[L_\tau(Y, f_\theta(X))],
\]

where the quantile loss function \( L_\tau \) is defined as:

\[
L_\tau(y, q) = \begin{cases} 
\tau(y - q), & y > q, \\
(1 - \tau)(q - y), & y \leq q.
\end{cases}
\]

Given a dataset \( \{(x_i, y_i)\}_{i=1}^N \), the estimation of \( f_\theta \) is done through minimizing the empirical loss: \( \min_\theta \frac{1}{N} \sum_{i=1}^N L_\tau(y_i, f_\theta(x_i)) \). After solving this minimization problem to obtain the optimal \( \theta^* \), the conditional quantile of \( Y \) given \( X = x \) can be predicted by \( f_{\theta^*}(x) \).

The above procedure is for a fixed \( \tau \) only, or, for the conditional quantile of \( Y \) at a fixed probability level only. This local information of the conditional distribution \( p(y|x) \) is not sufficient when we aim to have an estimation of the full distribution. A direct solution is to estimate different quantiles for different probability levels \( \tau_1 < \tau_2 < \cdots < \tau_k \) simultaneously and adopt an interpolation method to obtain the full distribution. Generally in simultaneous...
quantile regression, a K-dimensional parameterized mapping \( f_0(x) \) is specified and its k-th element \([f_0(x)]_k\) represents the conditional \( \tau_k \)-quantile of \( Y \). Then, the optimal \( \theta^* \) is found through solving the following optimization problem:

\[
\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} I_{\tau_k}(y_i, [f_0(x_i)]_k).
\]  

(3)

However, after this, we may find that the resulting quantiles \( f_0'(x) \) may cross, i.e., \([f_0'(x)]_j > [f_0'(x)]_k\) for some \( x, j, \) and \( k \) satisfying \( j < k \). This violates the basic principle that a cumulative distribution function and its inverse, i.e., the quantile function, should be monotonically increasing. This may occur because we do not impose any restrictions on the elements of \( f_0(x) \) in the optimization. An easy solution is rearranging the order of \([f_0'(x)]_1, \ldots, [f_0'(x)]_K\) such that they are monotonically increasing [7]. This is how we tackle this issue in this paper, although there are also other solutions [37] [4] which are more sophisticated.

2.2 Factor Copula Model

Probability integral transform is a way to convert any continuous random variable following a given distribution to a uniformly distributed random variable. This enables the development of marginal distribution modeling and dependence modeling among continuous random variables \( Y_1, Y_2, \ldots, Y_n \), which is called the copula approach. Specifically, suppose \( Y_i \) has cumulative function \( F_i \), then \( U_i := F_i(Y_i) \sim U(0,1) \). One can define a copula function \( C(u_1, u_2, \ldots, u_n) \) to describe the joint cumulative function of \( U_1, U_2, \ldots, U_n \) instead of modeling \( Y_1, Y_2, \ldots, Y_n \) directly. This is particularly useful when every \( Y_i \) has different and complex marginal distributions.

In high-dimensional cases, it is very often that all \( U_i \) are further converted to some commonly used random variables such as Gaussian or Student’s t: \( V_i := G^{-1}(U_i) \), where \( G \) is the cumulative function of the new random variable specified. Taken Gaussian as an example, now \( V_i \) is standard normally distributed and thus we can define dependence structure among \( V_1, \ldots, V_n \) easily by a full correlation matrix. However, it is more practical to assume a one-factor structure [9] among them in high-dimensional cases:

\[
V_i = \beta_i M + \sqrt{1 - \beta_i^2} Z_i,
\]  

(4)

where both \( M \) and \( Z_i \) are assumed to be standard normal variables. \( M \) is a common factor shared by all \( V_i \), and \( Z_i \) is the idiosyncratic variable. \( \beta_i \) measures the sensitivity of \( V_i \) to \( M \), thus incurs dependence between \( V_i \) and \( V_j \) when \( \beta_i > 0, \beta_j > 0, i \neq j \). The correlation between \( V_i \) and \( V_j \) is \( \beta_i \beta_j \).

2.3 R-copula

In most cases the random variable we model is continuous, thus the ordinary copula approach can be applied. However, in some special cases, the variable is the mixed type of discrete and continuous. The repayment rate of non-performing consumer credit is a typical example and is the focus of this paper. Loan borrowers may repay zero amount with a large probability or repay a continuous amount between zero and the sum of principal and interest. Now the probability integral transform cannot be applied directly. We address this issue with the so-called R-copula approach.

If the cumulative function \( F \) of the variable \( Y \) is not continuous at \( y_0 \), there must be \( F_+(y_0) - F_-(y_0) > 0 \). Thus the probability integral transform \( F(Y) \) will not take value in the interval \((F_-(y_0), F_+(y_0))\). \( F(Y) \) is not standard uniform anymore. To fix it, R-copula [12], [34] proposed to add another uniformly distributed variable to fill up this interval when \( y = y_0 \): \( F_+(y_0) + U'(F_+(y_0) - F_-(y_0)) \), where \( U' \sim U(0,1) \) is independent. The formal definition of this transform called random probability integral transform is

\[
U := F_-(Y) + U'(F_+(Y) - F_-(Y)).
\]  

(5)

When \( Y \) takes value at a continuous point, \( F_+(Y) = F_-(Y) \) and \( U = F(Y) \) as usual. At a discontinuous point, the standard uniformly distributed \( U' \) will fill up the gap between \( F_+(Y) \) and \( F_-(Y) \). When there are only finite discontinuous points, it is not hard to prove that \( U \) follows a standard uniform distribution. Moreover, the inverse transform from \( U \) to \( Y \) exists and is unique, making this method useful in practice. When we aim to model the dependence among \( Y_1, \ldots, Y_n \), the new variables \( U_1, \ldots, U_n \) can be constructed through this transform, and the Gaussian copula approach can be applied as usual. However, this method adds additional randomness caused by \( U' \) into \( U \), which needs careful treatment by us in real applications.

3 THE METHOD

In this section, we first describe the problem setup, and then elaborate on the architecture of the method we propose. Our final goal is to predict the distribution of a non-performing consumer credit portfolio’s overall repayment rate, as a further input of a pricing formula for risky assets. Our method architecture includes the quantile regression for marginal repayment rate prediction of individual loans, the R-copula approach for the mixed discrete and continuous rates, and the factor copula model for dependence modeling among loans. After fitting all unknown parameters, we simulate many repayment rates for all loans from our method, and further approximate the distribution of the portfolio’s overall repayment rate. The entire workflow is also illustrated in Figure 1.

3.1 Problem Setup

Suppose at time \( t \), we have \( N_t \) loans in the pool that have defaulted. Typically \( t \) denotes at which month we are inspecting. Furthermore, these \( N_t \) loans are divided into \( j \) categories according to their overdue stages, e.g., 30+ days, 60+ days, etc. The \( j \)-th category has \( N_{jt} \) loans included and their repayment rates in a fixed future time horizon are denoted as \( R_{ijt} \), where \( i \in I_{jt} \) and \( I_{jt} \) is the corresponding index set. The repayment rate is defined as

\[
R_{ijt} = \frac{\text{actual repayment amount}}{\text{expected repayment amount}}. 
\]  

(6)

We denote the expected repayment amount (generally it is the sum of principal and interest) in the fixed future time horizon as \( P_{ijt} \). So \( R_{ijt}P_{ijt} \) is the actual repayment amount. Typically this future time horizon can be chosen to be one year for the need of pricing.

For this portfolio of the \( N_t \) loans, the overall repayment rate is

\[
R_t = \frac{\sum_{j=1}^{J} \sum_{i \in I_{jt}} R_{ijt}P_{ijt}}{\sum_{j=1}^{J} \sum_{i \in I_{jt}} P_{ijt}}. 
\]  

(7)
We tackle this challenge with the proposed method containing the \( \tau \text{i} \) where \( \text{NN} \) is a neural network-based model with the parameter set \( \Theta \). The subscript \( i \in \mathcal{I}_j \) should be modeled and estimated. This includes the marginal distribution of every \( R_{ijt} \) and the dependence structure among them. These rates are not independent, because the credit status of a large population may be driven by some common underlying risk factors, such as the business cycle or the exogenous shock like COVID-19. In no situation this task will be easy due to the heterogeneity of every loan borrower’s credit behavior and the high-dimensional property when considering a huge-size portfolio. We tackle this challenge with the proposed method containing the two core parts described in Section 3.2 and 3.3 respectively.

### 3.2 Marginal Distribution Prediction

The first component of our proposed method is predicting the marginal distribution of \( R_{ijt} \), whose cumulative function is denoted as \( F_{ijt} \). To improve the model flexibility, we do not assume a parametric form for \( F_{ijt} \), but instead approximate it non-parametrically by the quantiles of a lot of probability levels spreading all over \((0, 1)\). Suppose the probability levels we choose are \( \tau_1 < \cdots < \tau_K \), the \( \tau_k \)-quantile of \( F_{ijt} \) is denoted as \( q_{ijt}^k \). Furthermore, suppose we have the feature vector \( x_{ijt} \) for each loan, we use the following model to predict the quantiles of \( F_{ijt} \):

\[
[q_{ijt}^1, \ldots, q_{ijt}^K]^T = \Psi(x_{ijt} | \Theta_j),
\]

where \( \Psi(x|\Theta) \) is a neural network-based model with the parameter set \( \Theta \). The subscript \( j \) in \( \Theta_j \) means we set different models for different categories of loans. \( \Psi(x|\Theta) \) is specified as following:

\[
\Psi(x|\Theta) = \text{clip(\text{linear(NN}_L(x)))},
\]

where \( \text{NN}_L \) is a \( L \)-layer ordinary neural network with the sigmoid activation function. The output of \( \text{NN}_L \) is then fed to a linear layer and the following clip operator which makes the final output be between the possible minimum and maximum values of \( R_{ijt} \).

To train the model for every category, or to learn the parameter set \( \Theta_j \) for a fixed \( j \), the observations \( r_{ijt} \) of \( R_{ijt} \) across different \( t \) and different \( i \in \mathcal{I}_j \) are collected to solve the following quantile loss minimization problem:

\[
\min_{\Theta_j} \frac{1}{\sum_j N_{jt}} \frac{1}{K} \sum_{t} \sum_{i \in \mathcal{I}_j} \sum_{k=1}^K L_{\tau_k} \{r_{ijt}, [\Psi(x_{ijt}|\Theta_j)]_k\},
\]

where \([ \cdot ]_k\) denotes the \( k \)-th element of a vector. After the optimal \( \Theta_j \) is obtained, the quantiles \( q_{ijt}^1, \ldots, q_{ijt}^K \) can be calculated through \( \Psi(x_{ijt}|\Theta_j) \) and a sorting operation is applied on them to ensure the monotony property, as described in Section 2.1. After that, a linear interpolation among these quantiles is applied to approximate the quantile function of \( R_{ijt} \), i.e., the inverse of \( F_{ijt} \). Thus \( F_{ijt} \) is also approximated, for further use in the following steps of our method.

### 3.3 Multi-name Dependence Modeling

Default dependence among loans is crucial for pricing any credit portfolio. Joint defaults occur more frequently than in the independent case. To model it, we first handle the issue of mixed discrete and continuous variables with R-copula, then use Gaussian copula to model the dependence among all loans’ repayment rates.

For a fixed \( t \), suppose we have obtained every \( F_{ijt} \) from the previous step. In the scenario of non-performing credit pricing, \( F_{ijt} \) is discontinuous at \( r = 0 \) and continuous otherwise. There is a jump from zero to a positive value at \( r = 0 \), which means borrowers may repay zero amount with a positive probability. Hence, we apply the R-copula approach described in the section of preliminaries and obtain the standard uniformly distributed \( U_{ijt} \) after transforming \( R_{ijt} \). Next, the one-factor Gaussian copula is applied as usual:

\[
V_{ijt} = G^{-1}(U_{ijt}),
\]

\[
V_{ijt} = \beta_{ij} M_t + \sqrt{1 - \beta_{ij}^2} Z_{ijt},
\]

where \( G \) is the cumulative function of standard Gaussian. Now \( U_{ijt} \sim N(0, 1) \), and both the latent \( M_t \) \sim \( N(0, 1) \), \( Z_{ijt} \sim N(0, 1) \). The subscript \( t \) in \( M_t \) means all loans in the pool in a fixed time share the same common factor. A large negative value of \( M_t \) will cause all \( V_{ijt} \) tend to have negative values at time \( t \), thus small \( U_{ijt} \) and netage \( R_{ijt} \) are more possible. The subscript \( jt \) in the parameter \( \beta_{ij} \) means the loans in the same category share the same sensitivity to the common factor. This is called the homogeneous credit portfolio assumption which is used for simplifying the model.
3.4 Dependence Parameter Estimation

\( \beta_{jt} \) is a measure of the dependence among all loans in a category in a fixed time. A large \( \beta_{jt} \) will induce more risk into the credit portfolio because borrowers tend to have lower repayment rates. In credit derivatives of corporate bonds, a large \( \beta \) means joint defaults will happen more likely. Thus, a precise estimate of this parameter is crucial for the pricing and risk management of credit derivatives. In credit derivatives of corporate bonds, \( \beta \) can be implied or recovered from the market quotes and the implied \( \beta \) can be used to price new credit derivatives. However, very differently, large pools of small consumer loans are not traded in a market in our situation. We have only one choice: estimating \( \beta \) from historical data.

The difficulty may come from the high-dimensional nature of the problem, i.e., typically a pool may have thousands or even millions of small consumer loans. Fortunately, our model framework is succinct enough to allow us to do the estimation. Suppose we observed the realizations of \( R_{ijt} \) as \( r_{ijt} \) and the corresponding realizations of \( V_{ijt} \) as \( u_{ijt} \) after the R-copula transform, then the realizations \( u_{ijt} \) of \( V_{ijt} \) can be obtained by Equation (11). From Equation (12), conditional on \( M_t = m_t \) in the fixed time \( t \) and when \( j \) is also fixed, all \( V_{ijt} \) are independently and identically distributed (i.i.d.). This inspires us to apply the generalized method of moments (GMM)[10] to estimate unknown parameters.

Specifically, when fixing \( t \) and \( j \), the conditional first and second moments of \( V_{ijt} \) are:

\[
E[V_{ijt}|M_t = m_t] = \beta_{jt} m_t, \quad (13)
\]

\[
\text{Var}[V_{ijt}|M_t = m_t] = 1 - \beta_{jt}^2, \quad (14)
\]

when \( m_t \) is also unknown, the above two equations have two unknown parameters \( m_t \) and \( \beta_{jt} \) which need estimation. Noticed that \( m_t \) is shared by every \( j \)-th category, separate estimation for different \( j \) will yield inconsistent values of \( m_t \). Thus, we combine these two equations for all \( j = 1, \ldots, J \). Now \( J + 1 \) unknown parameters are in \( 2J \) equations, making GMM applied successfully. Replacing the expectations with empirical sums, we have the final equations:

\[
\bar{\theta}_{jt} := \frac{1}{N_{jt}} \sum_{i \in I_{jt}} u_{ijt} = \beta_{jt} m_t, \quad j = 1, \ldots, J, \quad (15)
\]

\[
\text{Var}_{jt} := \frac{1}{N_{jt} - 1} \sum_{i \in I_{jt}} (u_{ijt} - \bar{\theta}_{jt})^2 = 1 - \beta_{jt}^2, \quad j = 1, \ldots, J. \quad (16)
\]

There may be no solutions to the above equation system, so we solve the squared loss minimization instead:

\[
\min_{m_t; \beta_{j1}, \ldots, \beta_{jJ}} \sum_{j=1}^{J} (\beta_{jt} m_t - \bar{\theta}_{jt})^2 + \sum_{j=1}^{J} (1 - \beta_{jt}^2 - \text{Var}_{jt})^2. \quad (17)
\]

Here we have finished the estimation for unknown parameters at time \( t \). This estimation is separate for different \( t \), such that we do not need long historical data. Actually, the cross-sectional information is rich enough for us to conduct such an estimation. At last, we will do this estimation many times and take an average of the estimated parameters to reduce the inaccuracy caused by additional randomness incurred by \( U^{*} \) in the R-copula approach.

3.5 Simulation for Portfolio

Our aim is to have an estimation or prediction of the full distribution of the entire portfolio’s repayment rate at every time \( t \), as defined in Equation (7). So far, we have prepared two components that are sufficient to achieve this goal. In Section 3.2, we have approached the marginal distribution of \( R_{ijt} \) \((j = 1, \ldots, J \) and \( i \in I_{jt} \) by a neural network-based model’s output. In Section 3.3 and 3.4, we modeled and estimated the dependence among \( R_{ijt} \) \((j = 1, \ldots, J \) and \( i \in I_{jt} \) through the copula approach. These two components are enough to determine the joint distribution of \( R_{ijt} \), thus are enough to determine the distribution of \( R_t \) too.

We are unable to derive the analytical form of the joint distribution of \( R_{ijt} \). Thus simulation method is adopted, and the workflow is the inverse of the modeling system. At every time \( t \), in every simulation step, we start from Equation (12) and simulate a realization from \( M_t \) and a realization from \( Z_{ijt} \) for every \( j \) and \( i \). Then a realization of \( V_{ijt} \) for every \( j \) and \( i \) is obtained by Equation (12). We then apply Equation (11) as well as the inverse of \( R \)-copula to obtain a realization of \( R_{ijt}, V_j = 1, \ldots, J \) and \( i \in I_{jt} \). By Equation (7), a realization of \( R_t \) is calculated. This process can be repeated many times for us to have an empirical estimation of the distribution of \( R_t \). A pricing formula will take this empirical distribution as an input, which is not the focus of our paper anymore.

4 EXPERIMENTS

In this section, we comprehensively describe the experiments we conduct. First, we collect consumer loan data from one of the biggest online platforms all over the world for credit lending to consumers. Then the proposed architecture is implemented on the large amount of non-performing loan data. We do fine analysis on the performance of our method, and also discuss its merits and limitations.

4.1 The Data

At the beginning of each month in the whole year of 2019, we consider it as the day of inspection and randomly select about 360,000 overdue loans from the central data warehouse to form the portfolio. According to their overdue stages, we split them into \( J = 17 \) categories named as \( M_1, M_2, \ldots, M_{12}, M_{13-15}, M_{16-18}, M_{19-21}, M_{22-24}, \) and \( M_{24+} \), where \( M_i \) means that the days overdue are more than \( i \) months but less than \( i + 1 \) months. There is a small chance that one loan may appear in different days of inspection, but in is different stages. Naturally, when \( i \) grows, the number of loans in \( M_i \) will decrease rapidly. To ensure the same stability of our model for every category, we select balanced numbers of loans for all overdue stages. For example, the number of loans in \( M_{1} \) at the date 2019-01-01 is about 10,000, and so is it for \( M_i, i = 2, \ldots, 36 \). We merge some stages to form the final \( J = 17 \) categories.

With every loan, we extract its features at the time of inspection as well as its label which is the repayment rate in future one year. The features are 84-dimensional covering comprehensive information about the loan and the borrower, such as the loan’s basic information, the borrower’s repayment history, the borrower’s searching and purchasing behaviors on the platform, and so on. The repayment rate of each loan and the overall repayment rate of a group of loans are calculated by Equation (6) and (7). The estimation and prediction of the latter is our final goal. Before starting, we
We first use the data collected from the date 2019-01-01 to 2019-10-31, where CL is the abbreviation for clip and linear operations, NN refers to the network layers. We have 10 time points and at each, a collection of samples is clipped to the range (0,1], which means the features are input into the linear and neural network layers. The linear layer has 99 output nodes and the output of samples follows a uniform distribution over (0, 1). Then, following the workflow of our method, the parameters of the dependence model are estimated. To reduce the error caused by extra randomness in R-copula, this estimation is performed ten times and the average estimated parameters are taken. At last, in the simulation for portfolio loss distribution, 1000 realizations of \( R_t \) are obtained to approximate the distribution of \( R_t \).

What we propose here is a bottom-up model, i.e., dealing with each individual loan first, followed by combining them to get the portfolio loss distribution. Another class of methods for credit portfolio modeling is called top-down, i.e., the portfolio loss distribution is specified directly without using the information of individual constituent loans [13]. To confirm the necessity of constructing a bottom-up model as we do, and also to demonstrate the merits of our model, we implement a benchmark top-down model for comparison which simply uses past four months’ moving average of each category’s overall repayment rates as the prediction of next repayment rate for that category. The prediction of the whole portfolio’s repayment rate is done by combining all categories together. However, this method cannot obtain the distribution of \( R_t \).

### 4.3 W-test

Through the simulation, we can predict the distribution of \( R_t \), which is denoted as \( F_t \). Ideally, the variable \( F_t(R_t) \) should follow a uniform distribution over (0, 1). So, if we have predicted \( F_t \) for \( R_t \), \( t \in T \), and if we also have the observed realization of \( R_t \), which is denoted as \( r_t \), we can test if the quantities \( F_t(r_t) \), \( t \in T \) follow a uniform distribution over (0, 1). Furthermore, this test is equivalent to testing if \( G^{-1}(F_t(r_t)) \) follow a standard normal distribution, where \( G \) is the cumulative function of standard normal. Therefore, finally we will do a normality test on \( G^{-1}(F_t(r_t)) \), \( t \in T \).

The number of our observation dates is 10 in the first testing set. So, Shapiro-Wilk test (W-test) can be used to test if a small set of samples follow the normal distribution [36]. Suppose the data samples we need to test are \( y_1 \leq y_2 \leq \cdots \leq y_{|T|} \) are the order statistics of \( y_i \) for \( t \in T \) (we denote \( n = |T| \) here), then the statistics of W-test is defined as:

\[
W = \frac{\left( \sum_{i=1}^{n} a_i y(i) \right)^2}{\sum_{i=1}^{n} y(i)^2},
\]

where the coefficients \( a_1, a_2, \cdots, a_n \) have specific values when the sample size is \( n \). The null hypothesis of the test is that the data follows the normal distribution. This hypothesis is rejected if \( W \) is below a critical value under a certain confidence level, say 95% chosen by us in this paper. A p-value close to 1 is an indicator of good performance, and a \( p \)-value less than 0.05 is a strong indicator of rejection of the null hypothesis.

### 4.4 Results

We first examine the predicted distribution of \( R_t \) obtained by our method on the testing set, through plotting the histogram of the simulated realizations of \( R_t \). Figure 3 shows the histograms given by the four versions of our method: CL-FC, CLNN2-FC, CLNN4-FC, and CLNN6-FC.
Figure 3: The empirical distribution of $R_t$, or, the histogram of the simulation of $R_t$. The yellow line and dark blue lines are the mean and mean±std of the simulation respectively. The red line is the true observed value of $R_t$. Here we only show the results of the dates 2019-01-01, 2019-06-01, 2019-11-01, and 2019-12-01 from top to down. Each column represents one version of our method: CL-FC, CLNN$_2$-FC, CLNN$_4$-FC, and CLNN$_6$-FC. For the limit of pages, we only place the histograms of the portfolios at four dates: 2019-01-01, 2019-06-01, 2019-11-01, and 2019-12-01. The latter two have no overlap with the dates of the training set. In Figure 3, the mean and mean±std of the simulation

| Date      | CL-FC         | CLNN$_2$-FC   | CLNN$_4$-FC   | CLNN$_6$-FC   |
|-----------|---------------|---------------|---------------|---------------|
|           | Truth | Mean  | Std   | Skewness | Kurtosis | Truth | Mean  | Std   | Skewness | Kurtosis |
| 2019-01-01| 0.0758 | 0.0815 | 0.0136 | 0.8852   | 1.4365   | 0.0758 | 0.0725 | 0.0099 | 0.5158   | 0.3759   |
| 2019-06-01| 0.0749 | 0.0814 | 0.0127 | 1.2853   | 3.8079   | 0.0749 | 0.0727 | 0.0083 | 0.4934   | 0.9094   |
| 2019-11-01| 0.0832 | 0.0885 | 0.0102 | 0.6639   | 0.8275   | 0.0832 | 0.0740 | 0.0077 | 0.2699   | 0.1589   |
| 2019-12-01| 0.0819 | 0.0898 | 0.0107 | 0.6926   | 0.9156   | 0.0819 | 0.0744 | 0.0076 | 0.5126   | 0.6511   |

Table 1: Related statistics of the predicted distribution of $R_t$. and CLNN$_6$-FC. For the limit of pages, we only place the histograms of the portfolios at four dates: 2019-01-01, 2019-06-01, 2019-11-01, and 2019-12-01. The latter two have no overlap with the dates of the training set. In Figure 3, the mean and mean±std of the simulation
of each histogram, as well as the true observed repayment rate of \( R_t \), are also plotted. We can see that in many cases, the true observation lies in an appropriate range of the histogram. We also list the higher-order moments of every predicted distribution of \( R_t \) in Table 1, which may be the extra inputs of a pricing formula.

To rigorously evaluate the performance of the distribution prediction, we conduct W-test for each version of our method over the first testing set which has 10 dates (W-test requires the dataset size \( n \geq 8 \)). All test results are reported in Table 2. As shown in Table 2, only linear regression-based method gets a significant rejection of the hypothesis that \( G^{-1}(F_t(r_t)), t \in T \) follow a standard Gaussian. In some sense, the performances of CLNN\(_2\)-FC and CLNN\(_4\)-FC are good enough because the W-test statistics are close to 1. Thus a mild layer size is enough in the specific design of our architecture. This may be because that the features we select as the input of the quantile regression have been proven previously to contain the most information relevant to the default prediction.

To further evaluate the performance of our method in a different way, we choose the mean absolute percentage error (MAPE) as another evaluation criteria. If we use \( \hat{r}_t \) to denote the mean of the predicted distribution of \( R_t \) (the yellow lines in Figure 3), then the criteria MAPE can be calculated as:

\[
\text{MAPE} = \frac{1}{|T|} \sum_{t \in T} \frac{|r_t - \hat{r}_t|}{r_t}.
\]

In MAPE, we take \( \hat{r}_t \) as the predicted value of \( r_t \), and it is consistent with the idea that the predicted value should be as close as possible to the true observed value \( r_t \). Since we have two testing sets, we report the results on both sets for comprehensive comparisons. We also include the performance of the top-down method we implement into comparisons.

Table 3 presents the MAPE values of different methods on the first testing set covering 10 months. Not surprisingly, the neural network-based methods work much better than linear regression-based method. Among the neural network-based methods, CLNN\(_4\)-FC yields the best performance. More crucially, the top-down method performs much worse than our methods do, indicating the absolute necessity of constructing a bottom-up system for the pricing of non-performing consumer credit portfolios, as we do in this paper.

Table 4 presents the MAPE results over the second testing set covering the non-overlap two months. Different from the first MAPE table, CLNN\(_6\)-FC works best in the non-overlap case. Nevertheless, the significance of this phenomenon may be restricted to the sample size. We could not have more data points over time because constructing such a dataset needs a observation period of future one year. So we should have records of the loans up to the end of the year 2020, which is unavailable when this research is done.

To conclude all the experiment results, the method we construct is able to capture both the marginal distribution of every loan and the dependence structure among loans in a portfolio, and consequently can yield good prediction of the distribution of overall repayment rate of a credit portfolio containing non-performing consumer loans. Thus this workflow or architecture has great potential of being successfully applied to the real business domain.

4.5 Drawbacks and Future Works
As one can find, the methods or techniques we adopt here are mainly traditional statistical ones, except that in individual loan’s repayment forecasting we use neural networks. This is primarily because AI currently lacks (popular) methodologies for: i) effective probabilistic forecasting (as opposed to point forecasting in supervised machine learning), which is of great need in finance; ii) the dependence modeling of extremely high-dimensional variables. For the second one, [41] recently proposed a generative machine learning model for capturing tail dependence structure among stock returns. However, it remains a quite challenging task and needs more further research works.

Another potential drawback of our approach is the homogeneous assumption in the dependence modeling, i.e., the constant \( \beta \) across a category of loans. Although it is a common setting in credit portfolio pricing frameworks for simplifying, it may miss the possible more complex dependence structure in the data. Future works can aim at alleviating this drawback in a more data-driven manner.

5 CONCLUSIONS
We have described in detail our bottom-up system for the overdue loan portfolios’ repayment risk prediction. Every part of our architecture is elaborately designed and finally, the experiments demonstrate the success of our system. It can be used in the pricing

---

### Table 2: The W-test results. The CL-FC method yields a significant rejection. CLNN\(_2\)-FC and CLNN\(_4\)-FC produce W-test statistics that are close to 1.

| Method   | W-test statistics | p-value |
|----------|------------------|---------|
| CL-FC    | 0.7916           | 0.0114  |
| CLNN\(_2\)-FC | 0.9759      | 0.9397  |
| CLNN\(_4\)-FC | 0.9798      | 0.9641  |
| CLNN\(_6\)-FC | 0.9602      | 0.7886  |

### Table 3: The MAPE results. CLNN\(_4\)-FC gives the best performance while the top-down method is the worst.

| Method   | MAPE  |
|----------|-------|
| CL-FC    | 9.36% |
| CLNN\(_2\)-FC | 4.10%  |
| CLNN\(_4\)-FC | 3.38%  |
| CLNN\(_6\)-FC | 4.20%  |
| top-down | 10.81%|

### Table 4: The MAPE results over the non-overlapping testing set. CLNN\(_6\)-FC gives the best performance while the top-down method is the worst.

| Method   | MAPE (non-overlap) |
|----------|---------------------|
| CL-FC    | 8.00%               |
| CLNN\(_2\)-FC | 10.10%   |
| CLNN\(_4\)-FC | 9.09%    |
| CLNN\(_6\)-FC | 4.42%    |
| top-down | 14.09%             |
and risk management of non-performing consumer credit portfolios and may benefit both the new-emerging risk transferring business and the ordinary risk management tasks.

In this paper, we do not explicitly give the price of a credit portfolio. What we offer is the expected profit or cash flow (the mean repayment) generated by holding the portfolio, as well as the uncertainty measures such as volatility (standard deviation), skewness, and kurtosis. One can then determine his/her own acceptable price to adjust scorecards. Journal of the Operational Research Society 66, 10 (2015), 1656–1668.

Yanjie Kang, Rumbang Cui, Jiang Deng, and Ning Jia. 2019. A novel credit scoring framework for auto loan using an unbalanced-learning-based reject inference. In 2019 IEEE Conference on Computational Intelligence for Financial Engineering & Economics (CIFEr), IEEE, 1–8.

Ralf Kellner, Maximilian Nagl, and Daniel Rösch. 2021. Opening the Black Box: Quantile Neural Networks for Loss Given Default Prediction. Journal of Banking & Finance (2021), 106334.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In International Conference on Learning Representations, ICLR 2015.

Roger Koenker and Kevin F Hallock. 2001. Quantile regression. Journal of economic perspectives 15, 4 (2001), 143–156.

Jean-Paul Laurent and Jon Gregory. 2005. Basket default swaps, CDOs and factor copulas. Journal of risk 7, 4 (2005), 103–122.

David X Li. 2000. On default correlation: A copula function approach. The Journal of Fixed Income 9, 4 (2000), 43–54.

Wei Li, Shuai Ding, Yi Chen, Hao Wang, and Shulin Yang. 2019. Transfer learning-based default prediction model for consumer credit in China. The Journal of Supercomputing 75, 2 (2019), 862–884.

Carlos Madeira. 2019. Measuring the covariation risk of consumer debt portfolios. Journal of Economic Dynamics and Control 104 (2019), 21–38.

Rashmi Malhotra and Davinder K Malhotra. 2003. Evaluating consumer loans using neural networks. Omega 31, 2 (2003), 83–96.

Nicola Meinshausen and Greg Ridgeway. 2006. Quantile regression forests. Journal of Machine Learning Research 7, 6 (2006).

Dominic O’Kane. 2011. Modelling single-name and multi-name credit derivatives. Vol. 573. John Wiley & Sons.

Vincenzo Pacelli, Michele Azzollini, et al. 2011. An artificial neural network approach for credit risk management. Journal of Intelligent Learning Systems and Applications 3, 02 (2011), 103.

Xiaohua Peng and Steven Kou. 2009. Default clustering and valuation of collateralized debt obligations. Working Paper (2009).

Daniel Rösch and Harald Schurel. 2004. Forecasting retail portfolio credit risk. The Journal of Risk Finance (2004).

Ludger Rüschendorf. 2009. On the distributional transform, Sklar’s theorem, and the empirical copula process. Journal of statistical planning and inference 139, 11 (2009), 3921–3927.

Hyun-Sun Ryu. 2018. Understanding benefit and risk framework of fintech adoption: Comparison of early adopters and late adopters. In Proceedings of the 51st Hawaii International Conference on System Sciences.

Samuel Sanford Shapiro and Martin B Wilk. 1965. An analysis of variance test for normality (complete samples). Biometrika 52, 3/4 (1965), 591–611.

Ichiro Takeuchi, Quoc Le, Timothy Sears, Alexander Smola, et al. 2006. Nonparametric quantile estimation. (2006).

Lyn Thomas, Jonathan Crock, and David Edelman. 2017. Credit scoring and its applications. Advances in Neural Information Processing Systems.

Dezhong Wang, Svetlozar T Rachev, and Frank J Fabozzi. 2009. Pricing tranches of a CDO and a CDS index: Recent advances and future research. Risk Assessment (2009), 283–286.

Qi Wu and Xing Yan. 2019. Capturing deep tail risk via sequential learning of quantile dynamics. Journal of Economic Dynamics and Control 109 (2019), 103717.

Xing Yan, Qi Wu, and Wen Zhang. 2019. Cross-sectional learning of extremal dependence among financial assets. Advances in Neural Information Processing Systems (2019), 3852–3862.