MS&E 246 Financial Risk Analytics Final Report

Yifei Guo[†], Yiting Zhao[‡],Zhengji Yang[‡]

Abstract. In this project, we designed models to predict the small-business loan default and the loss given default based on a data set of roughly 150,000 equipment loans backed the US Small Business Administration (SBA) between 1990 and 2014. Based upon preliminary findings from a naive proportional hazard model, we first built a time-varying cox model, tailored to our dataset with right censoring and ties, to predict the default probabilities. Non-linearity, time-varying effects and data stratification are all incorporated into our model. The default model works well in predicting defaults in loans, achieving an in-sample and out-of-sample AUC of 0.726 and 0.724. Then, a loss model is built to predict the loss at default which ensembles a random forest classifier to predict whether the loan will fully recover a random forest regressor to predict the loss percentage. The ensemble model can reduce RMSE from 0.2439 to 0.2217. To testify our model performance, we construct a portfolio backed by 500 small-business loans and predict the pool-level default rate and loss amount in 5 and 10 years. As our time-varying cox model requires quaterly macroeconomic data during the sample period, we develop a data augmentation method to generate required data for our prediction based on the real statistics in 2005-2014. Then we pass the simulated data into our fitted cox model and random forest regressor to predict default events and corresponding loss amount. Finally, we use this simulation to estimate loss distribution for the selected portfolios to understand risk and help risk management.

1. Introduction

Risk management has long been an important area to study both in academia and in industry. Its importance was especially emphasized after the financial crisis of 2007-2008 when the complexity of the interconnectivity underlying the financial system was revealed. One default would trigger chained effects across multiple institutes via financial derivatives. On the one hand, loans are important to energize business activities within an economy, sustain economic growth and facilitate innovative ideas and products. On the other hand, however, excess loans and a high leverage ratio within a society would trigger financial crises due to defaults. Analyzing default events and predicting default losses are indispensable for building a healthy financial system.

In this project, we focus a dataset on the small business loans backed by the US Small Business Administration between 1990 and 2014. SBA rarely makes direct loans but facilitates and backs loans given to small businesses by lending partners (SBA, 2023). Because lending to small businesses is, in general, riskier than lending to large corporations, small businesses, while usually are the ones who desperately need financial supports especially during crises, tend to be left behind by the economy. By providing guarantees to small business loans, SBA effectively reduces risks faced by the lenders while providing small business with easier access to capital. In addition to the SBA dataset, we also try to include other macroeconomics factors that relate to the loans default.

As the cox model, originally developed for medical areas, was developed specifically to model the occurrence of default-like events, we focus primarily on using the cox model for default predictions. As the economic conditions are always changing and play an important role in triggering default events, a time-varying cox model without assuming proportionality of hazard is more suitable for this dataset. To be robust, preliminary model is

[†] ICME, Stanford University; guoyifei@stanford.edu.
[‡] MS&E, Sanford University; yitzhao@stanford.edu.
[†] ICME, Stanford University; yangzj@stanford.edu.

built to motivates the final model build-up.

In addition to predicting the default, we also try to predict the loss given default. We utilize random forest to explore non-linearity and intersections across features and leverage ensemble to further improve the prediction accuracy. Given that some loans could fully recover and yield a loss of 0, we first predict whether a loan will fully recover given default with a random forest classifier, then predict the loss if the loan will not fully recover with a random forest regressor.

Finally, we select 500 loans to form a portfolio and try to understand the investment risk behind it. As our model requires quarterly macroeconomic data during the simulation period, i.e. 10 years, we develop a data augmentation method to simulate them. We choose real macroeconomic statistics during 2005 and 2014, then recursively generate quarterly data by adding a perturbed real data difference to simulated data of last quarter, which is

$$Y_{t+1} = Y_t + (1 + r_{long\,term} + r_{long\,term} * r_t) * (X_{t+1} - X_t).$$

Here, $r_{long\,term}$ factor is deterministic for each simulation, describing the long term macroeconomic situation. While r_t factor depicts the short term volatility for each quarter. $\{X_t\}$ are the real macroeconomic data and $\{Y_t\}$ are the simulated data. Using this mechanism, we simulate the future macroeconomics condition for the selected portfolio. We examine the risk of the selected portfolio in 5 years and 10 years from perspective of loss distribution, Value at Risk, Average Value at Risk, [5%,15%] tranche and [15%,100%] tranche. This simulation gives us some intuition about the risk management.

This report is organized as the following. Section 2 explores the SBA dataset. Section 3 introduces preliminary findings that motivates our final model. Section 4 explains thoroughly our default model build-up. Section 5 analyzes the default model results and performs validation checks. Section 6 introduces the workflow of how we predict the loss given default and presents the model performance. Section 7 describes how to simulate for future macroeconomics and how we analyze the risk of the selected portfolio. Section 8 briefly talks about the future work. Section 9 concludes all the key findings.

2. Exploratory Data Analysis

This section includes an overview of the SBA dataset, data cleaning, 500-loan portfolio selection, and visualizations of the data trend. The goal of this section is trying to understand the data and to inform our model building.

2.1. SBA Dataset Overview

We have two main tasks in our project. The first is to predict the default probability and the second is to predict the loss given default. The original SBA dataset contains 147,423 loans and 30 columns. To improve data quality, we have excluded loans that are canceled or without loan status, those with 0 terms and those in states other than U.S. 50 mainland states. Moreover, we have removed columns that are irrelevant to our goal, such as names of borrowers, projects and lenders. To avoid overfitting, we extract the first 3 digits of zipcode to inform the address information and extract the first 2 digits of the Naics Code to capture the industry information. This leaves us with 126,282 loans and 13 columns.

Table 1 summarizes the SBA dataset after data cleaning and some feature engineering.

Column Name	Description
Gross Approval	Total Loan Amount
Gross ChargeOff Amount	Total Loss; 0 means the loan was fully recovered.
Term In Months	Length of loan term
Project State	State where project occurs
Business Type	Borrower Business Type - Individual, Partnership, or Corporation
Loan Status	Current status of loan - PIF, CHGOFF or EXEMPT
Loan Purpose	Take keyword from subprogram; Categorical variable with 4 levels - Delta Loans, Premier Certified Lender Program, Private Sector Financed or Refinanced
is Same Borr Project	Indicator variable of whether the borrower is in the same state as the project
is Same Borr CDC	Indicator variable of whether the borrower is in the same state as the CDC
ApprovalDate	Date the loan was approved
End Date	Min(Charge off Date, Approval Date + Terms, 2014-01-31)
sub Zipcode	First 3 digits of borrower zipcode
sub NaicsCode	First 2 digits of NaicsCode

Table 1. SBA Dataset Overview

We also find a few issues that inform our model setup during the exploration. First, there are 8,865 defaults in the dataset and only 1,341 of them are fully recovered, thus there might be data imbalance issues when we are predicting the loss given default. Second, there are 6,039 loans with the same end dates, which leads to the concern of ties when we build the cox model. Third, 72,014 loans have Exempt status, indicating right censoring issues. We will try to deal with those issues when we build the model.

2.2. Portfolio Selection

For simulation purposes, we randomly select 500 loans that were approved after 2010 to form our 500-loan portfolio as we are trying to use loans that are most up-to-date. We will not use these 500 loans to fit the default model and the loss model.

2.3. Trend Visualizations

By exploring the dataset, we can find in figure 1, in general, default rates spiked for the loans that are approved around the financial crisis (2007-2009). This is what we can expect from our knowledge.

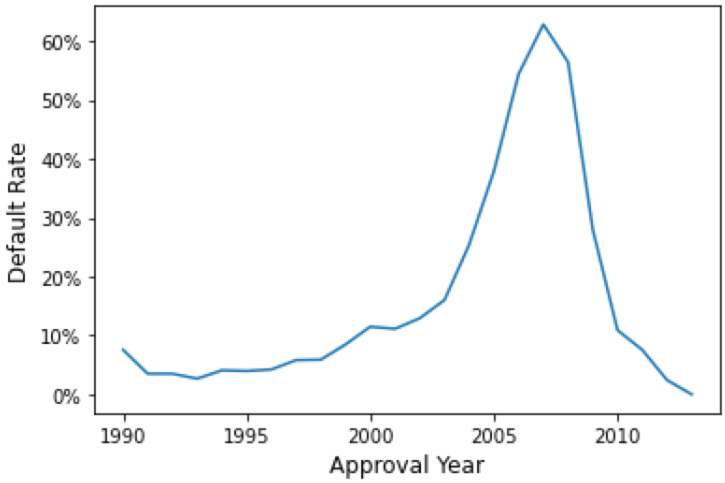


Figure 1. Default Rate by Approval Year

In addition, we also try to find other relevant factors to the prediction task. From figure 2, we can examine the relationship between default rate and business type by approval year. As shown in figure 2, we can observe different business types were affected differently. The different trajectories of the 3 curves imply the "Individual" Business Type suffered greater default rates than corporations and partnerships. Although corporations constitute a greater share of the data set, they exhibit medium default risk, as compared to the other business types. Taken together, this plot reveals business types were affected differently by the recession, offering useful signals for subsequent modeling.

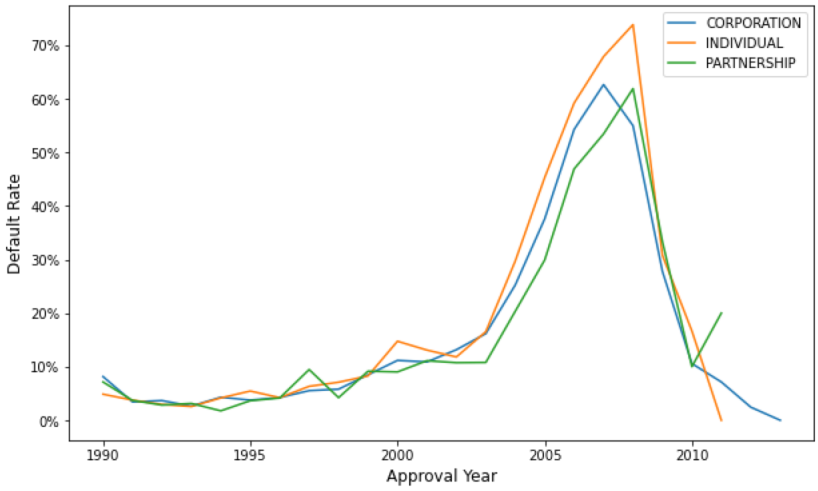


Figure 2. Default Rate v.s. Business Type by Approval Year

Next, we examined whether we would observe a similar time-dependent interaction effect between default rate and Loan Amount. The figure 3 below reveals that loans of all sizes that were approved around the Great Recession faced the greatest default rates. However, loans of sizes and $1m-$2m appear to have experienced larger default rates over time compared to other loans. Since loans of different sizes have different default rate

patterns over time, we would also expect the Loan Amount feature to offer predictive power.

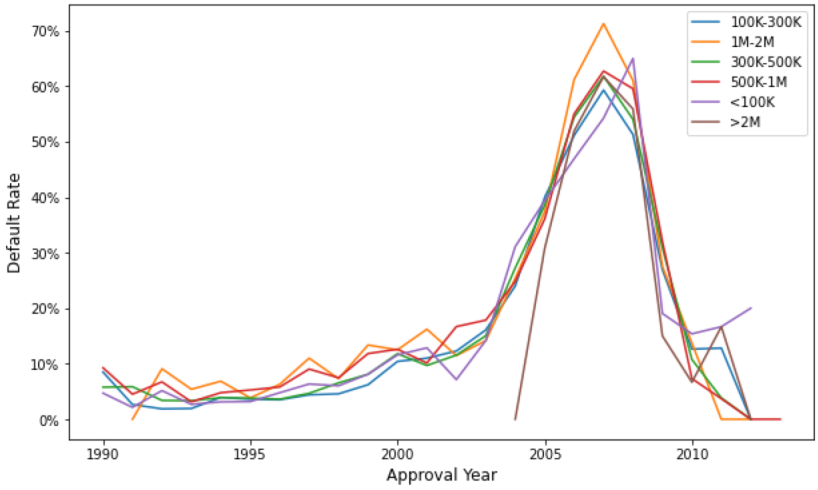


Figure 3. Default Rate v.s. Approval Amount by Approval Year

We also expect different economic sectors would exhibit different default rates over time. As the original North American Industry Classification System (NAICS) code for each loan is too detailed, we truncated it to the first two digits, which represent broad industry classes. Figure 4 shows the default rate for loans of each truncated NAICS code approved in each year between 1990-2014. We observe a considerable variance in default rates between sectors. For example, "Management of Companies and Enterprises" (code: 55) has one of the highest default rates during the recession. However, "Transportation and Warehousing" (code: 48) consistently have the lowest default rates. These patterns are consistent with our intuition and show the importance to include the truncated NAICS code.

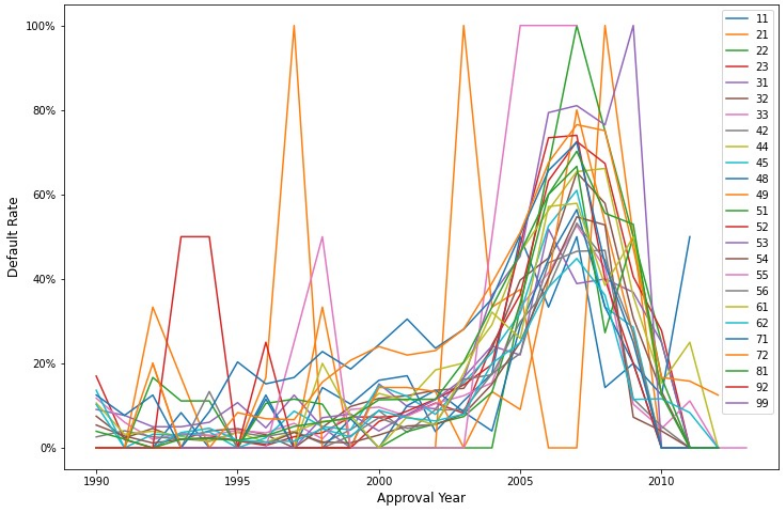


Figure 4. Default Rate v.s. Naics Code by Approval Year

We also compared the default rates for different loan purposes. Figure 5 below shows the default rates of the

different loan purposes. We observe that the loans under Premier Certified Lender Program and Private Sector Financed are most common and have higher default risk. On the other hand, loans for Delta and Refinance purposes are uncommon and have low default risk. This suggests loan purpose offers useful signal for predicting default risk.

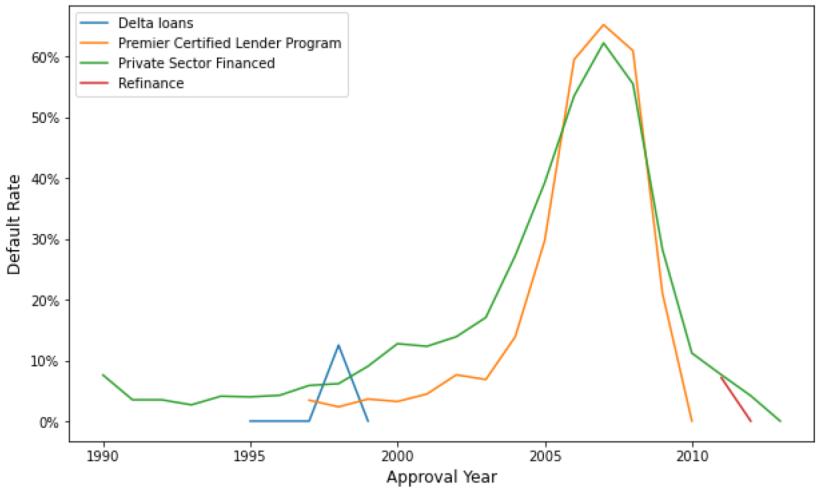


Figure 5. Default Rate v.s. Loan Purpose by Approval Year

Figure 6 examines the default rates for projects in different states. We observe although loans in the West typically have a lower default rate, they suffer a lot during the recession. In addition, loans in the South typically have a higher default rate and it holds true during the recession. Overall, projects in different states were affected differently, which can improve our prediction.

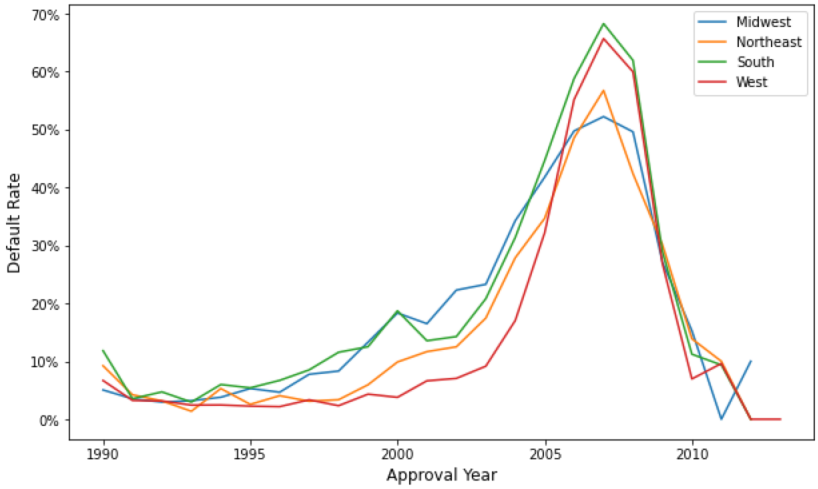


Figure 6. Default Rate v.s. Project State by Approval Year

The left-hand side of the figure 7 examines the default rate against whether the borrower is in the same state as the lender, while the right-hand side of the figure 7 examines the default rate against whether the borrower is in

the same state as the project. In short, we can see those not in the same state have higher default risks.

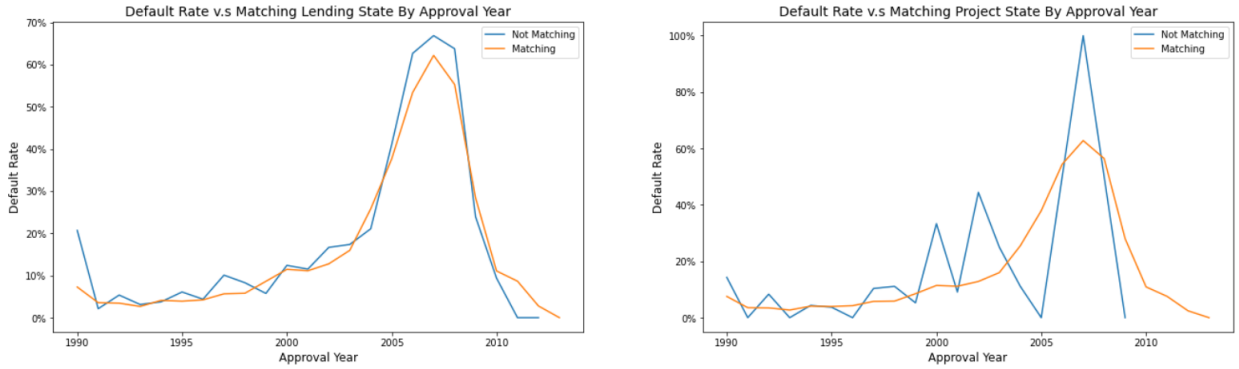


Figure 7. Default Rate v.s. Matching State by Approval Year

Figure 8 examines the relationship between default rates and loan terms. We can observe that most loans approved after 2000 are those between 72 month and 240 month. Additionally, loans with terms between 120 months and 240 month have a higher default rate, indicating the loans with a longer terms have higher default risk.

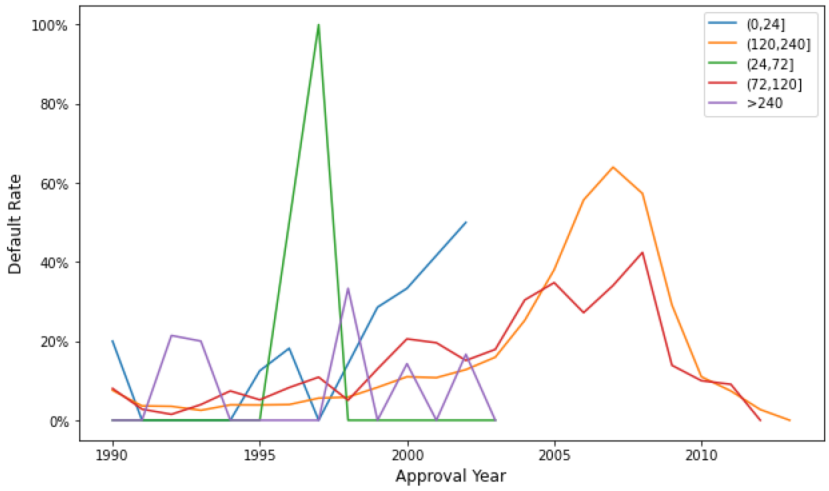


Figure 8. Default Rate v.s. Terms by Approval Year

3. Preliminary Findings

Before we built our default model, we first tried a naive approach using a proportional cox model, whose results inform us of the potential problems and the extra information awaiting to be considered. This section includes, an introduction of all the extra data used in both the preliminary model and in our final model, the preliminary model build-up, interpretations of the results, and analyses of the problems revealed.

Treated as Loan Feature (country level)	
log_S&P	The log of S&P 500 at the end date (default/censored/PIF)
VIX	Market VIX at the end date (market volatility)
TED	3M-LIBOR - 3M T-bill (credit risk in the economy)
PRIME	US Prime rate at the loan origination date
Leverage	Common equity tier 1 capital ratio: a bank's core equity capital / total risk-weighted assets

Table 2. Country-level external data

Treated as Time Varying Co-variates (State / zipcode level)	
UnemploymentRate	state-level unemployment rate
log_GSP	log of Gross State Product
log_HPI	zipcode-level House Price Index (HPI) from the Federal Housing Finance Agency

Table 3. Zipcode and state-level external data

3.1. External Data

Intuitively, macroeconomic situations play an important role in risk analytics. Thus, in addition to loan features included in the original dataset, we also include macroeconomic data in our model. The Table 2 includes all variables taken into account in the preliminary model. The variables names are pretty self-explanatory. We include the log_S&P to inform overall stock market condition, i.e. if we are in a booming or a contracting market. We took log of the S&P index to transform it into linear growth. VIX is included as information on market volatility. We also include TED which equals the difference between the three-month-Libor rate and the three-month-T-bill rate, representing the credit risk in the economy. Because interest rate data is not included in the original dataset, the U.S. prime rate is also considered as a baseline interest rate. The common equity tier 1 capital ratio is used to indicate the leverage condition of the financial system. Because we do not consider time varying effects in the preliminary model, these data are included as loan features chosen at specific times. TED and the U.S. prime rate are specified at the loan's origination date while all the other three variables are specified at the loan's end date (default / PIF / censored).

Data in table 3 are zipcode and state level data used in our final default model. Because for a fixed point in time, all the country-level data (variables in Table 2) are the same for all loans, they cannot be included in the time-varying model. The time-varying cox model incorporates them into the baseline hazard. Thus, we can only use zipcode and state level data as time-varying variables. We include the state unemployment rate, log_GSP (log of Gross State Product) and log_HPI (log of zipcode-level House Price Index). We take log on GSP and HPI to make conform with linearity.

3.2. Preliminary Model - Time Invariant Proportional Cox Model

As mentioned above, we fit a time-invariant cox model as our preliminary model with all of the loan feature variables and country-level economic variables included. Then the time invariant cox model can be depicted as

$$\lambda(x;\beta) = exp(\sum_{i=1}^{p} \beta_i x_i)$$

$$PL(\beta) = \prod_{i:\tau_i \leq T_i} \frac{\lambda(X_{i,\tau_i};\beta)}{\sum_{j:a=\tau_i - t_i} \lambda(X_{j,\tau_i};\beta)}$$

where X is the concatenate of loan feature variables and country-level macroeconomic variables; $\lambda(x;\beta)$ is the proportional hazard rate and $PL(\beta)$ is the partial likelihood defined by the product of conditional probability P_i, which is the probability of loan i default conditional on a loan default at time τ_i.

3.3. Preliminary Coefficients

Figure 9 is the fitting result of our preliminary model using the training dataset. This result reveals that all of the loan feature co-variates are significant: log_amount, TermInMonths, BusinessType, is_Same_Borr, is_Same_Borr_Project, except loan_purpose, which is not significant at 0.1 significance level for all categories. Thus, for computational efficiency, the loan purpose variable is dropped in our final default mode. Country-level variables are all significant at 0.01 significance level so we definitely need to incorporate these economic information into out default model. As mentioned before, because country-level data are the same for all loans at one specific time, directly including them into our time-varying model will trigger perfect co-linearity problem. Thus, state-level and zipcode-level alternatives are considered instead. Figure 10 is the fit of the average survival curve for the preliminary model on the training dataset.

```
Call:
coxph(formula = Surv(time, status == 2) ~ log_amount + TED +
    TermInMonths + log_amount + log_SP + VIX + PRIME + Leverage +
    BusinessType + is_Same_Borr_CDC + is_Same_Borr_Project +
    loan_purpose, data = dfori, ties = "efron", id = id)

                                            coef exp(coef)  se(coef)       z        p
log_amount                              0.350893  1.420336  0.012276  28.583  < 2e-16
TED                                     0.425410  1.530218  0.010017  42.468  < 2e-16
TermInMonths                            0.430135  1.537465  0.009474  45.403  < 2e-16
log_SP                                 -0.673034  0.510158  0.007603 -88.517  < 2e-16
VIX                                    -0.138417  0.870735  0.005052 -27.399  < 2e-16
PRIME                                  -0.688514  0.502322  0.014352 -47.974  < 2e-16
Leverage                               -0.534678  0.585858  0.009299 -57.497  < 2e-16
BusinessTypeINDIVIDUAL                 -0.105494  0.899880  0.036917  -2.858  0.00427
BusinessTypePARTNERSHIP                -0.530693  0.588197  0.061289  -8.659  < 2e-16
is_Same_Borr_CDCTRUE                   -0.514031  0.598080  0.038929 -13.204  < 2e-16
is_Same_Borr_ProjectTRUE                0.922236  2.514907  0.202655   4.551 5.35e-06
loan_purposePremier Certified Lender Program  1.428666  4.173128  1.000748   1.428  0.15341
loan_purposePrivate Sector Financed     0.780298  2.182123  1.000173   0.780  0.43529
loan_purposeRefinance                   0.826094  2.284379  1.416023   0.583  0.55963

Likelihood ratio test=29748  on 14 df, p=< 2.2e-16
n= 125707, number of events= 8852
```

Figure 9. Results from the preliminary Model

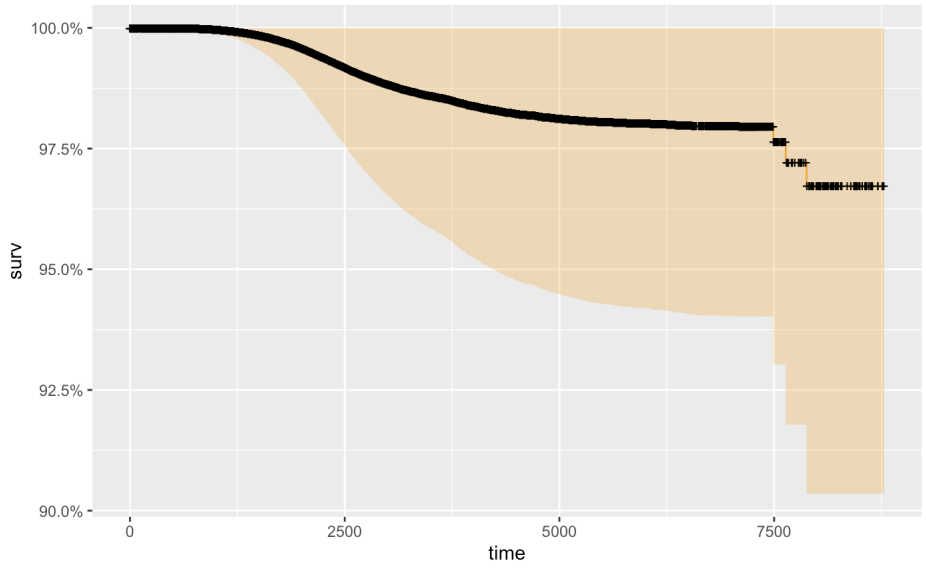


Figure 10. Results from the preliminary Model

3.4. Test Proportional Hazard Hypothesis

Our preliminary model also allows us to test if the proportional hazard hypothesis holds for our dataset. The following outcome (Figure 11) is computed using cox.zph test from the survival package. This test is based on weighted Score-process residuals and detailed explanations of the test can be found in this paper. (Grambsch and Therneau, 1994). A number of the variables and the Global result are significant, indicating that the proportional hazard assumption is violated.

	chisq	df	p
log_amount	7.67e+01	1	< 2e-16
TED	4.00e+02	1	< 2e-16
TermInMonths	1.25e+01	1	0.00041
log_SP	4.22e+01	1	8.3e-11
VIX	1.09e+02	1	< 2e-16
PRIME	6.09e+02	1	< 2e-16
Leverage	1.36e+00	1	0.24303
BusinessType	6.98e-01	2	0.70526
is_Same_Borr_CDC	1.19e+01	1	0.00057
is_Same_Borr_Project	5.07e-02	1	0.82189
loan_purpose	7.56e+00	3	0.05615
GLOBAL	1.56e+03	14	< 2e-16

Figure 11. Results from cox.zph()

Given that the proportional hazard hypothesis is violated, what are the remedies we can consider? Grambsch and Therneau in their book *Model survival data: extending the Cox model* suggest four ways to remedy the unproportionality (Grambsch and Therneau, 2000):

1. Incorporate co-variates as stratification factors
2. Partition the time axis

3. Time-dependent co-variates or time-dependent coefficient: $\beta(t)X = \beta X(t)$

4. Different model

Based on the availability and the characteristics of our dataset, we implement a combination of stratification and time-varying model by incorporating zipcode and country-level macroeconomic variables as time-varying co-variates.

3.5. Test for Linearity

The preliminary model also allows us to check if any of the variables exhibits non-linearity. Grambsch and Therneau (Grambsch and Therneau, 2000) suggests the following algorithm:

1. Estimated a null model with only the intercept term.

```
Call:  coxph(formula = Surv(time, status == 2) ~ 1, data = dfori)

Null model
  log likelihood= -99543.58
  n= 125782
```

2. Calculate martingale residuals of the proportional hazard model.
   - Martingale Residual is defined as $MR_i = \delta_i - r_{ci}$, where r_{ci} is based on the Nelson Aalen estimate of the cumulative hazard function (pp.115). (Collet, 2014)
   - Properties of Martingale Residual:
     - $\mathbf{E}(MR_i) = 0$, $\sum \hat{M}_i = 0$, $cov(M_i, M_j) = 0$, $cov(\hat{M}_i, \hat{M}_j) < 0$ (Grambsch and Therneau, 2000).
     - By definition Martingale residuals $\in [1, -\infty]$ for uncensored observations and $\in [0, -\infty]$ for censored observations.
     - Interpretation: indicating whether or not a loan defaulted as expected by the model.
     - Interpretation: a residual closer to the upper limit of a martingale residual is obtained when a loan as an unexpectedly short default time.

3. Plot the martingale residuals against the co-variates.

4. Analyze the plots.
   - The leftmost plot in the first line, Figure 12, has time as the x-axis. As the martingale residuals have a downward-curved pattern, time dependence is revealed. Outliers present but because we are working with default with limited default events, we chose not to remove these outliers. Alternatively, we can use deviance residuals, corrected for symmetry, to identify residuals. The results are similar.
   - The martingale residuals for **TermInMonths** exhibit non-linear pattern, revealed by the leftmost plot in the second line, Figure 13.
   - No clear non-linearity for other variables.
   - Because all censored loans end at the last date and some economic variables are specified as the value at the end date, the martingale residuals for censored data are sometimes clustered.

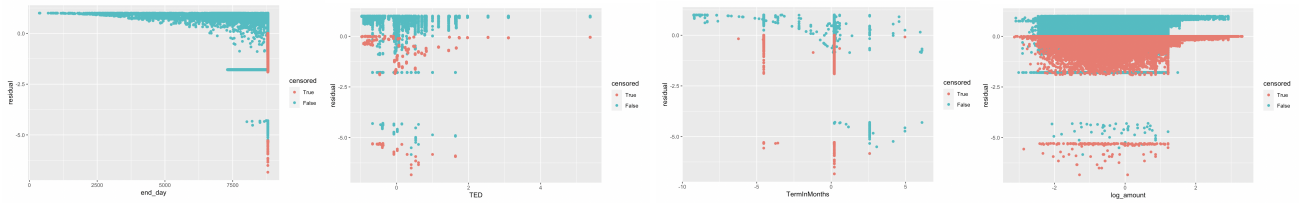


Figure 12. Martingale Residual Plots

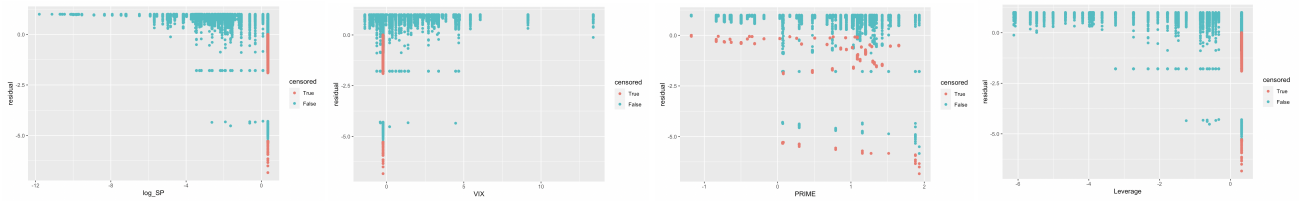


Figure 13. Martingale Residual Plots

Summing up, we need to include loan features except loan_purpose; add state and zipcode level economic data as time varying co-variates; add nonlinear term for TermInMonths and consider right-censoring and ties.

4. Default Model Build-Up

Based on preliminary results, our default model build-up is introduced in this section, which is organized as the following. Section 4.1 is a brief recap of the time-varying cox model. Section 4.2 incorporates right-censoring into the time-varying model. Section 4.3 incorporates ties into the time-varying model. Section 4.4 describes how we impose non-linearity structure on TermInMonths and the penalty used to better fit the model. Section 4.5 formulates how time-varying dataset is constructed. Section 4.6 briefly introduces the fitting algorithm used.

4.1. Time-Varying Cox Model Recap

Let T_i^* denote the time from loan origination to default. We assume that T_i^* are iid with pdf $f(t)$, (Assumption 1). Let $S_t = P(T^* > t)$ denotes the survival function. Then the hazard function is

$$\lambda(t) = lim_{h \to 0} P(t \leq T^* < t + h \mid T^* \geq t)/h = \frac{f(t)}{S(t)}$$

For each loan i, the cox model formulates the hazard as $\lambda_i(t) = \lambda_0(t) e^{X_i(t)\beta}$

Because the baseline hazard $\lambda_0(t)$ is the same across all loans, the hazard ratio our main focus:

$$\frac{\lambda_i(t)}{\lambda_j(t)} = \frac{\lambda_0(t) e^{X_i(t)\beta}}{\lambda_0(t) e^{X_j(t)\beta}} = \frac{e^{X_i(t)\beta}}{e^{X_j(t)\beta}}$$

Thus, the cox model is indeed a model for relative risk. A discussion of our next step to build an additional model for this baseline hazard in included in Section 8.1.

4.2. Right-censoring

In our dataset, for each loan there are three types of end state: except for an observed PIF or default event, the loan may be still alive at the end of the sample period, for which we would use right censored technique to deal with. Let C_i^* denotes the time from origination to the end of the dataset, i.e. the censoring time. Let $T_i = min(T_i^*, C_i^*)$ denotes the followup time. Here we make the non-informative assumption that assumes C_i^* is independent of T_i^* (Assumption 2). Then, the likelihood with censoring is:

$$\begin{aligned} &\prod_{i:\, default} f(T_i^*|X_i) \times \prod_{j:\, PIF} F(T_j^*|X_j) \times \prod_{k:\, censor} F(C^*|X_k) \\ =& \prod_{i:\, default} \lambda(T_i|X_i) \times \prod_{j=1}^{n} F(T_j|X_j) \end{aligned}$$

where $F(t|X_j) = F(\tau \leq t - t_j|X_j;\beta) = exp(-\int_{t=t_j}^{t} \lambda(X_j,\beta)ds)$. For comparison, without censoring for events i = 1 ... n happened at T_i^*, the likelihood is $\prod_{i:\, default} f(T_i^*|X_i) \times \prod_{j:\, PIF} F(T_j^*|X_j) = \prod_{j:\, default} \lambda(T_j^*) \prod_{i=1...n} F(T_i^*)$. In other word, right censoring technique treats those exempt loans as paid in full at censoring time.

4.3. Ties

There are a total of 6039 loans with tied end dates (PIF / Default / Censoted). Due to the presence of ties in our dataset, we cannot assume that each event time corresponds to exactly one event. To understand the effect of ties, let's first define $Y_i(t)$ as an indicator of whether unit i is under observation and at risk at time t. As we are going to use the Survival Package in R to fit the model, we follow its formulation of partial likelihood as in the original paper written by Cox (Cox, 1972):

$$PL(\beta) = \prod_{i=1}^{n} \prod_{t \geq 0} \frac{Y_i(t)exp[X_i(t)\beta]}{\sum_j Y_j(t)exp[X_j(t)\beta]} \tag{1}$$

When two events E_1 and E_2 happen at the same time. How we order E_1 and E_2 affects the partial likelihood formulated as Eq. 1. Hence, we need to use an approximation scheme to adjust the partial likelihood to fit this setting. Grambsch and Therneau (Grambsch and Therneau, 2000) propose four different approximation schemes:

4.3.1. Breslow approximation

This is the simplest approximation proposed independently by Breslow and by Peto, which uses the complete summation for every denominator in tied events, i.e.

$$PL_{Breslow}(t) = \prod_{i=1}^{d} \frac{\lambda_i}{\sum_{k=1}^{d} \lambda_k + \sum_{j:\, T_j^* > t} \lambda_j},$$

However, it is the least accurate approximation. In particular, it counts failed individuals more than once in the denominater, producing a conservative bias, hence estimating β too close to 0.

4.3.2. EFRON APPROXIMATION

As a better approximation, Efron approximation uses average denominator, i.e.

$$PL_{Efron}(t) = \prod_{i=1}^{d} \frac{\lambda_i}{(d-i+1)/d \cdot \sum_{k=1}^{d} \lambda_k + \sum_{j:\, T_j^* > t} \lambda_j},$$

in which d is the number of tied defaults at time t.

A possible interpretation of this method is that each of the (tied) default loans is certain to be in the first denominator, has $(d-1)/d$ chance of being in the second,..., $(d-i)/d$ chance of being in the $i+1$-th denominator. If the tied defaults all happen to have an identical risk score, then the Efron solution will be exact.

4.3.3. EXACT PARTIAL LIKELIHOOD

The exact partial likelihood for d tied defaults out of n loans at risk is

$$PL_{exact}(t) = \frac{\prod_{i=1}^{d} \lambda_i/(1+\lambda_i)}{\sum_{S(d,n)} \prod_{j=1}^{d} \lambda_{k_j}/(1+\lambda_{k_j})}$$

From expression of the exact partial likelihood, the computational cost is $O(n^d)$, but Gail et al. develop a recursive algorithm to accelerate it greatly. But for large sample number n and large tied events number d, it is still an expensive method.

4.3.4. AVERAGE LIKELIHOOD

DeLong et al. have shown that the averaged partial likelihood has an equivalent representation in terms of integrals, which is

$$PL_{average}(t) = \int_0^{\infty} \prod_{l=1}^{d} [1 - exp(\frac{r_l t}{\sum_{k=d+1}^{n} r_k})] e^{-t} dt$$

Actually, these integrals are well-conditioned numerically, with positive smooth integrands, hence can be efficiently evaluated by numerical approximation methods. The resulting algorithm is $O(d^2)$, far faster than the nominal $O(d!)$.

We select the Efron approximation for this project because it balances complexity and efficiency. Although simple, the Breslow approximation is the least accurate one among the four. Efron approximation builds upon the Breslow approximation and manages to have simple computations. The Exact Partial Likelihood conducts an exhaustive enumeration but changes the model's functional form to a logistic model:

$$\frac{\lambda_i(t)}{1+\lambda_i(t)} = \frac{\lambda_0(t)}{1+\lambda_0(t)} exp(X_i(t)\beta)$$

which disturbs our interpretation of the model. Averaged Likelihood approach also involves an exhaustive enumeration. Even it has numerical evaluation by using integrals, it's still a costly method. Hence, the Efron approximation is the best fit here.

4.4. Non-linearity and Penalty (Grambsch and Therneau, 2000)

As discussed in the preliminary results, TermInMonth exhibits nonlinear behavior. To model the nonlinear effect of it on the hazard rate, we use Penalized Smoothing Splines (psplines) for prediction. The smoothing spline (of some order) is

$$f(x) = \sum_{i=1}^{N} \theta_i \cdot B_i(x)$$

where $\{B_i(\cdot)\}_{i=1}^{N}$ is a truncated B-spline basis (of some order) and $\theta_i \in \mathbb{R}$ are the corresponding coefficients. This approach is flexible because by choosing proper N, $f(\cdot)$ can be a good approximation to any function. This approach is also robust because with local support on $B_i(\cdot)$, $f(\cdot)$ will not diverge to infinity beyond the sample range. As the spline can be quite complex and may overfit, a roughness penalty $J(f)$ is added into the objective

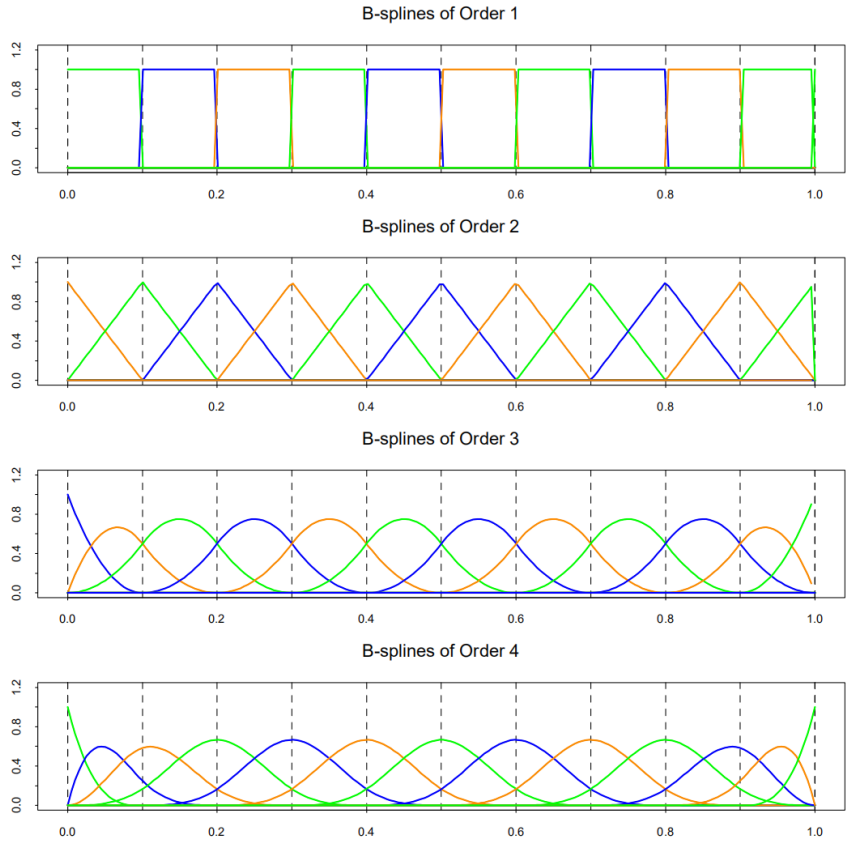


Figure 14. B-Spline Basis

function. Hence for hazard model $\lambda(t) = exp(X(t)\beta + \sum_{j=1}^{m} f_j(Z_j(t)))$, the loss function is (we don't consider the penalty for β here)

$$\mathcal{L}(\beta, \{\theta_j\}, \{X_i\}) = log(PL(\beta, \{\theta_j\}, \{X_i\})) + \tilde{\lambda_2} \sum_{j=1}^{m} J(f_j)$$

$$J(f_j) = \int ||{f_j}^{(2)}(x)||_2^2 dx = \sum_{l=1}^{N}\sum_{k=1}^{N} \theta_l^j \theta_k^j \int B_l^{(2)}(x) B_k^{(2)}(x) dx = {\theta^j}^T \mathcal{K}_f \theta^j.$$

where $f_j^{(2)}$ is the second order derivatives of f_j. If we let $T \in \mathbb{R}^{N-2\times N}$ be the matrix of second differences, i.e. $T_{i,i} = 1$; $T_{i,i+1} = -2$; $T_{i,i+2} = 1$ and $T_{i,j} = 0$ for other entries.

4.5. Time Varying Co-variates

In order to incorporate the time-varying features of the covariates, we sliced our dataset into time-interval slices according to (Therneau et al., 2023). Figure 15 is a snapshot of our final dataset. The following steps are performed:

- for loan i originated at t_{i1} and ends at t_{i2}, we cut the time interval between t_1 and t_2 into disjoint intervals with the length of each intervals equal to, a quarter, the time between t_{i1} and the start of the next quarter after origination or the time between t_{i2} and the end of the previous quarter before the end data.
- variable time_start is added to indicate the start of the time intervals.
- variable time_end is added to indicate the end of the time intervals.
- variable time is added representing T_i.
- variable status indicates the final status of a loan. 0: censored, 1: PIF, 2: default.
- variable default indicates whether default happened in this interval.
- all time-varying co-variates are merged respect to each time interval and loan characteristics. clusterid is added to the model code to specify which loan does a row correspond to.

id	status	default	time_start	time_end	time	TermInMonths	log_PersonalIncome	UnemploymentRate	log_HPI	is_Same_Borr_Project	BusinessType	log_amount	log_GSP
0	1	0	1	90	365	12	10.71080	4.3	4.694371	TRUE	INDIVIDUAL	12.01974	11.72845
0	1	0	90	181	365	12	10.72652	4.3	4.694371	TRUE	INDIVIDUAL	12.01974	11.72845
0	1	0	181	273	365	12	10.74221	4.3	4.694371	TRUE	INDIVIDUAL	12.01974	11.72845
0	1	0	273	365	365	12	10.74400	4.3	4.694371	TRUE	INDIVIDUAL	12.01974	11.72845
0	1	0	365	366	365	12	10.74144	4.4	4.694371	TRUE	INDIVIDUAL	12.01974	11.72845
1	1	0	1	90	7305	240	13.35555	5.8	4.619862	TRUE	INDIVIDUAL	11.66993	14.45550
1	1	0	90	181	7305	240	13.37036	5.8	4.619862	TRUE	INDIVIDUAL	11.66993	14.45550
1	1	0	181	273	7305	240	13.38047	5.8	4.619862	TRUE	INDIVIDUAL	11.66993	14.45550
1	1	0	273	365	7305	240	13.39311	5.8	4.619862	TRUE	INDIVIDUAL	11.66993	14.45550
1	1	0	365	455	7305	240	13.39427	7.7	4.619862	TRUE	INDIVIDUAL	11.66993	14.45550
1	1	0	455	546	7305	240	13.40321	7.7	4.619862	TRUE	INDIVIDUAL	11.66993	14.45550
1	1	0	546	638	7305	240	13.41218	7.7	4.619862	TRUE	INDIVIDUAL	11.66993	14.45550
1	1	0	638	730	7305	240	13.42619	7.7	4.619862	TRUE	INDIVIDUAL	11.66993	14.45550
1	1	0	730	821	7305	240	13.44209	9.3	4.619862	TRUE	INDIVIDUAL	11.66993	14.45550
1	1	0	821	912	7305	240	13.45922	9.3	4.619862	TRUE	INDIVIDUAL	11.66993	14.45550
1	1	0	912	1004	7305	240	13.46976	9.3	4.619862	TRUE	INDIVIDUAL	11.66993	14.45550
1	1	0	1004	1096	7305	240	13.47788	9.3	4.619862	TRUE	INDIVIDUAL	11.66993	14.45550
1	1	0	1096	1186	7305	240	13.47910	9.5	4.619862	TRUE	INDIVIDUAL	11.66993	14.45550
1	1	0	1186	1277	7305	240	13.48439	9.5	4.619862	TRUE	INDIVIDUAL	11.66993	14.45550
1	1	0	1277	1369	7305	240	13.48637	9.5	4.619862	TRUE	INDIVIDUAL	11.66993	14.45550

Figure 15. First 20 lines of out dataset

4.6. Fitting Algorithm (Grambsch and Therneau, 2000)

As the default model portion is coded in R, the survival package is used to perform model fitting and validation. The survival package implemented the Newton-Raphson Algorithm to solve the partial likelihood equation. First, we start with an initial guess $\hat{\beta}^0$ and then compute $\hat{\beta}$ iteratively until convergence: $\hat{\beta^{n+1}} = \hat{\beta}^n + \mathcal{I}^{-1}(\hat{\beta}^n)U(\hat{\beta}^n)$.

5. Default Model Results and Validations

After building our default model, our final model, model results and validations are performd in this section. Section 5.1 introduces our final default model used. Section 5.2 presents the coefficient fitting for the time-varying cox model. Section 5.3 performs in-sample and out-of-sample validation of the model. Section 5.4 demonstrates the influences analyses.

5.1. Default Model

After incorporating time-varying covariates, nonlinear effect and tied defaults, our default model is:

$$\lambda(t, \beta, \{\theta^j\}, X^i, Z^i) = exp(X_i(t)\beta + \sum_{j=1}^{m} f_j(Z_j^i(t), \theta^j)).$$

in which X^i, Z^i are the linear and nonlinear features for i-th loan, $\lambda(t, \cdot)$ is the proportional hazard rate.

The loss function with penalty is:

$$\begin{aligned}\mathcal{L}(\beta, \{\theta^j\}|X, Z) = &\sum_{\tau}\sum_{j=1}^{d_\tau} log\frac{\lambda(\tau, X^j, Z^j)}{(d_\tau - i + 1)/d_\tau \cdot \sum_{i=1}^{d}{}_\tau \lambda(\tau, X^i, Z^i) + \sum_{k:\, T_k^* > \tau} \lambda(\tau, X^k, Z^k)} \\ &+ \mu_1||\beta||_2^2 + \mu_2 \sum_j \theta^{jT} \mathcal{K} \theta^j,\end{aligned}$$

5.2. Default Model Coefficient Fitting

Figure 16 is the fitting result for our default model using the training dataset. Here we include all variables from Table 3 and as much loan features as possible. We did not add all loan features because adding them all results in the model exceeding our computer memory capacity. We also try to add ridge panelty to as many variables as possible but due to the limited computational capacity, we only managed to add a ridge panelty for variable log_GSP. $H_0 : \beta_0 = ... = \beta_k = 0$ of the overall model is tested using the likelihood ratio test with statistics: $2 * l(\hat{\beta} - \hat{\beta}_0)$. The likelihood ratio test has p-value less than 2e-16 so we reject the H_0 hypothesis. The significance test for each coefficient is conducted using the Wald test with statistics: $(\hat{\beta} - \hat{\beta}_0)^T \hat{\mathcal{I}} (\hat{\beta} - \hat{\beta}_0)$, where $\hat{\mathcal{I}}$ is the information matrix.

The following is our interpretation of the coefficient β:

Insignificant Variables:

- The coefficient for **Business Type of Individual** is not significant, meaning that the default hazard for a loan issued for an individual business does not differ significantly from that of a corporate business. This seems to be counter intuitive as corporate business is usually more robust than individual business. However, because all loans are small business loans in our dataset, even corporate business can be small which may not differ much from individual business.
- The coefficient for **is_Same_Borr_Project** is not significant because it is very common to do business in a different location than the original location of the business.

- The coefficient for **log_GDP** is not significant. We find this result to be less intuitive as Gross State Product is an important indicator of the state's macro economic condition. A possible explanation is that GSP changes may be minor and lagged because GSP involves too many components. Thus, changes in GSP do not reflect whether the borrower or the project correspond to the loan is particularly affected by the economic condition.

Significant variables:

- $\hat{\beta_{TermInMonths}} = -0.16922$: Short-term loans have higher default hazard than long-term loans. Intuitively, short-term loans may be less robust to short-term fluctuations as small businesses may not be able to overcome short-term economics fluctuations. Thus, it is reasonable to expect a negative correspondence between TermInMonth and default hazard.

- $\hat{\beta_{UnemploymentRate}} = 0.0407$: When unemployment rate is higher, the macro-economy is more likely to be in a worse situation, resulting in higher default hazard.

- $\hat{\beta_{log\_HPI}} = -1.50746$: Because our data for HPI is on zipcode level with the base-year HPI set to 100, log_HPI, in fact, measure the relative change in housing price in a zipcode region relative to the base year. When log_HPI decreases in a region, its economic condition is very likely to be troublesome, contributing to higher default hazard. This result is consistent with the negative coefficient of log_HPI.

- $\hat{\beta_{Partnatship}} = -0.31097$: The default hazard for a loan issued for a partnership business is significantly lower than that of a corporate business. Businesses in partnership are generally based on a mature and market-proven business model so such business is more robust to uncertainties, resulting in less default hazard.

- $\hat{\beta_{log\_amount}} = 0.3182$: Larger loans face more uncertainties. If something unexpected happened but the loan amount is small, it is more likely that the borrower can still manage to pay back the loan whereas if the loan amount is large, borrowers face more challenges to pay back the loan in full.

- $\hat{\beta_{is\_Same\_Borr\_CDC}} = -0.27719$: We find this result to be counter-intuitive as if borrowers and CDC are in the same region, the local authority is more likely to provide help for the borrower. However, our modle suggests the opposite.

5.3. In-Sample and Out-of-Sample Validation

Figure 19 plots the ROC curve for both the testing and the training dataset. The area under the curve for the training dataset is 0.726 whereas the area under the curve for the testing dataset is 0.724. The ROC curve for the testing dataset is less smooth than that of the training dataset due to the fact that there are more data involved in the training set. In addition, both ROC curves have a very small portion below the 45 degree line. This indicates that our model does not predict defaults well on some extreme cases. We do not consider this as a problem of our model but due to some outlying data points.

```
Call:
coxph(formula = Surv(time_start, time_end, default == 2) ~ pspline(TermInMonths) +
    UnemploymentRate + log_HPI + is_Same_Borr_CDC + is_Same_Borr_Project +
    BusinessType + log_amount + ridge(log_GSP), data = df, x = TRUE,
    id = id, cluster = id)

                               coef  se(coef)       se2     Chisq DF       p
pspline(TermInMonths), li  -0.16922   0.00804   0.01944 442.74585  1 < 2e-16
pspline(TermInMonths), no                                241.52292  3 < 2e-16
UnemploymentRate            0.04070   0.00713   0.00738  32.55011  1 1.2e-08
log_HPI                    -1.50746   0.05869   0.06194 659.71323  1 < 2e-16
is_Same_Borr_CDCTRUE       -0.27719   0.03846   0.03916  51.95285  1 5.7e-13
is_Same_Borr_ProjectTRUE    0.24504   0.21424   0.21028   1.30819  1   0.253
BusinessTypeINDIVIDUAL      0.09291   0.03943   0.03904   5.55323  1   0.018
BusinessTypePARTNERSHIP    -0.31097   0.06474   0.06459  23.07026  1 1.6e-06
log_amount                  0.31820   0.01491   0.01478 455.75560  1 < 2e-16
ridge(log_GSP)             -0.01551   0.00657   0.00679   5.57121  1   0.018

Iterations: 5 outer, 18 Newton-Raphson
     Theta= 0.454
Degrees of freedom for terms= 4.0 0.9 1.0 1.0 1.0 2.0 1.0 0.5
Likelihood ratio test=1326  on 11.3 df, p=<2e-16
n= 4140476, number of events= 7949
```

Figure 16. Results from our default model

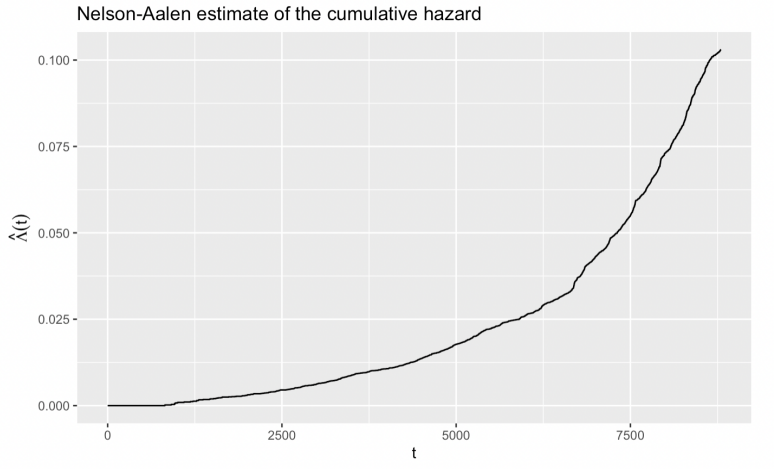


Figure 17. Nelson-Aalen estimate of the cumulative hazard

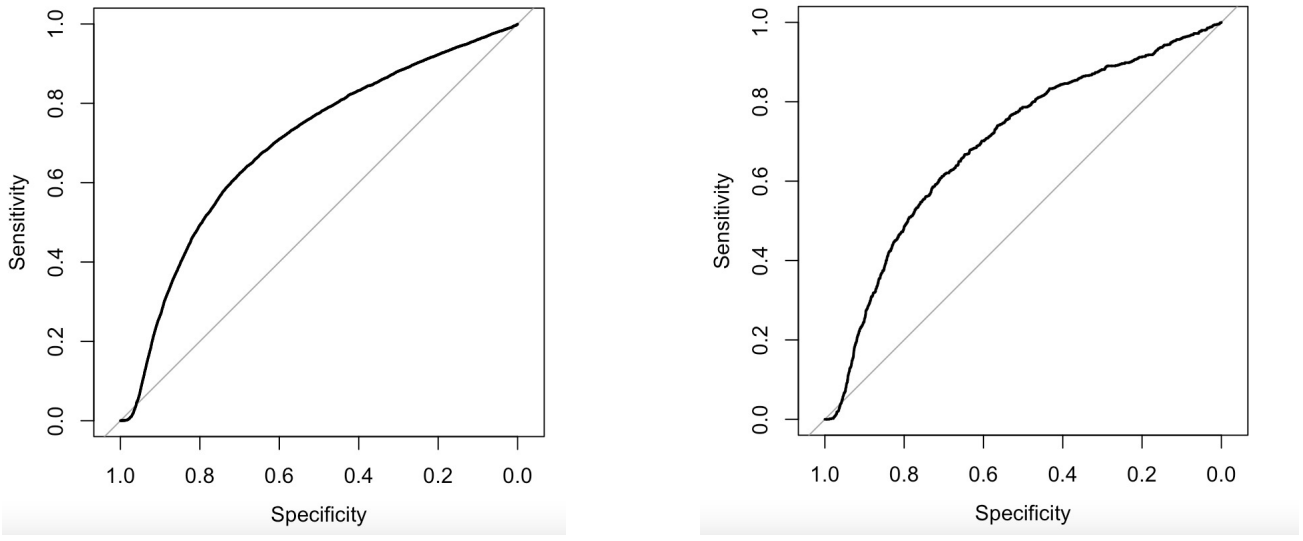


Figure 19. ROC Curve for training dataset and for testing dataset (AUC for Training: 0.726; AUC for Testing: 0.724)

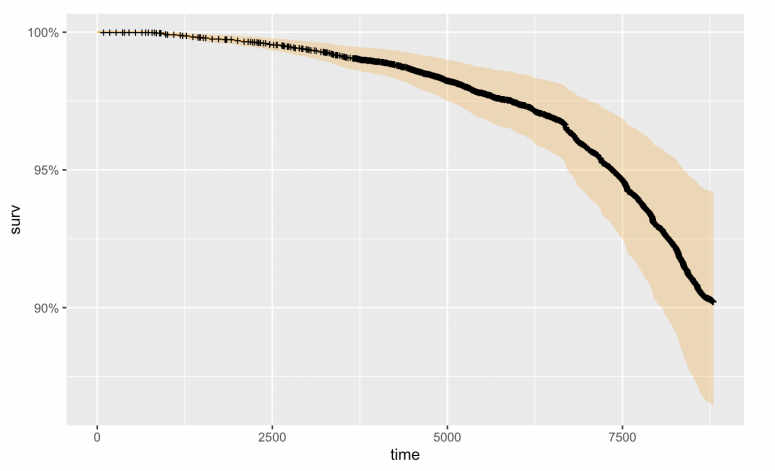
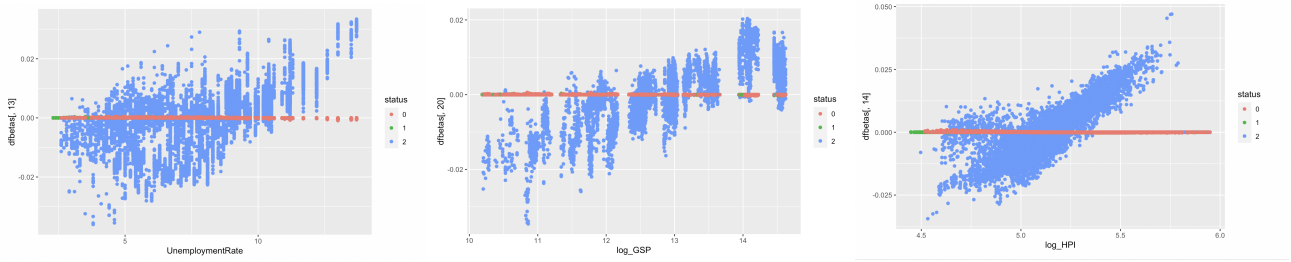


Figure 18. Estimated Survival Curve

5.4. Influences Analyses

We used $dfbetas_{ij} = \frac{\hat{\beta}_j - \hat{\beta_{(i)j}}}{SE(\hat{\beta}_j)}$ to assess the impact of each data point on the fit of our model. Figure 5.4 plots the dfbetas against the time-varying co-variates. From these plots, we can identify outliers from the these residual plots. The outliers are 5 defaulted corporate loans for project in DC with very large approval amount of more than $100000. We chose not to ignore them as we believe them are essential for risk management. Furthermore, there is a positive linear patterns exhibited in the dfbetas plot for variable log_HPI toward the right end. We think such nonlinear feature is likely due to some computational error in computation of Newton-Raphson Algorithm. The survival package has a caped number of iteration allowed so if more iterations are performed, our model results may be improved.



6. Loss Model

This section is about the model to predict loss given default. Section 6.1 introduces the data we used and the workflow of the model. Section 6.2 talks about the performance of our classifier. Section 6.3 talks about the performance of our ensemble model and compare it with another regressor.

6.1. Model Design

To predict the loss given default, we only use 8,865 defaults after excluding those in the 500-loans portfolio. There are two groups of features we have used. One is loan characteristics, including Gross Approval, Term in

Months, Project State, Business Type, Loan Purpose, is Same Borr Project, is Same Borr CDC, sub zipcode and sub NaicsCode. Another is macroeconomics, including S&P500, VIX, TED, PRIME, Leverage, state-level GSP, state-level Personal Income, state-level UnemploymentRate, Industry GDP, and zipcode-level HPI at the default.

Among 8,865 loans, we randomly select 1,000 loans to be our test set. We will train and tune hyper-parameters only on the train set.

Figure 20 summarizes how we ensemble two models to predict the loss. The machine learning model we used here is the random forest, as it can capture non-linearities and intersection effects among features, with generally better performance than linear models. Given that some loans are fully recovered and predicting them yields larger losses, we first train a random forest classifier to predict whether a loan will fully recover. If the classifier predicts the loan will fully recover, we will predict the loss to be 0, otherwise, we will use another random forest regressor to predict the loss given default. The target of the regression is the loss percentage rather than the loss amount, given that the variance of the loss amount may be too large and the loss amount relies heavily on the loan amount. After we have the loss percentage prediction, we will time it by gross approval amount to obtain the loss prediction.

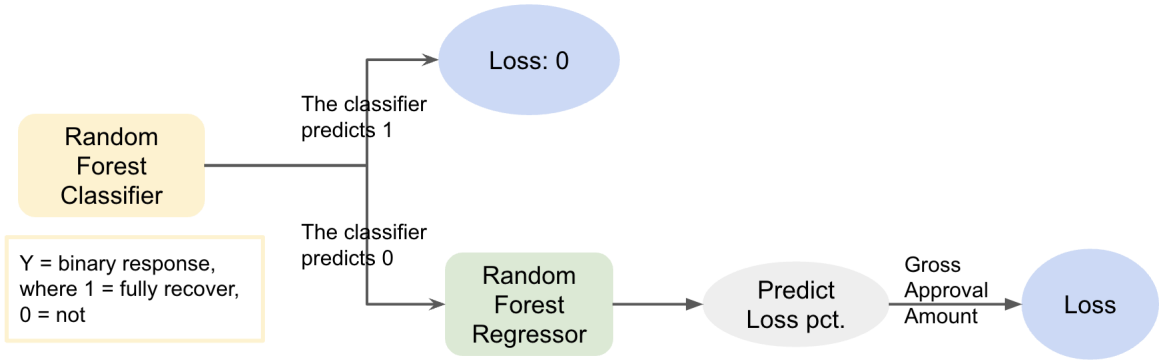


Figure 20. Workflow of Loss Prediction

6.2. Classifier Performance

The target of the random forest classifier is a binary variable, where 1 means fully recovered and 0 means not fully recovered. We will label our loans based on the charge-off amount. If the charge-off amount is 0, we will label the loan to be 1, to be 0 otherwise.

AS only 15% of the loans in the train set are fully recovered, we have a data imbalance issue here. Our approach to addressing imbalanced datasets is to resample the dataset, which combines oversampling the minority class and undersampling the majority class. The simplest oversampling approach involves duplicating examples in the minority class, although these examples don't add any new information to the model. Instead, new examples can be synthesized from the existing examples. This is a type of data augmentation for the minority class and is referred to as the Synthetic Minority Oversampling Technique, or SMOTE for short. SMOTE works by selecting

examples that are close in the feature space, drawing a line between the examples in the feature space and drawing a new sample at a point along that line. Specifically, a random example from the minority class is first chosen. Then k of the nearest neighbors for that example are found (typically k=5). A randomly selected neighbor is chosen and a synthetic example is created at a randomly selected point between the two examples in the feature space. Moreover, the original paper (Bowyer et al., 2011) on SMOTE suggested combining SMOTE with random undersampling of the majority class.

Binary classifiers are routinely evaluated with performance measures such as sensitivity and specificity, and performance is frequently illustrated with Receiver Operating Characteristics (ROC) plots. Many bioinformatics studies develop and evaluate classifiers that are to be applied to strongly imbalanced datasets in which the number of negatives outweighs the number of positives significantly. While ROC plots are visually appealing and provide an overview of a classifier's performance across a wide range of specificities, one can ask whether ROC plots could be misleading when applied in imbalanced classification scenarios. (Saito and Rehmsmeier, 2015) shows that the visual interpretability of ROC plots in the context of imbalanced datasets can be deceptive with respect to conclusions about the reliability of classification performance, owing to an intuitive but wrong interpretation of specificity. PRC plots, on the other hand, can provide the viewer with an accurate prediction of future classification performance due to the fact that they evaluate the fraction of true positives among positive predictions. Therefore, we use precision-recall curve and the area under precision-recall curve here ton evaluate the binary classifier.

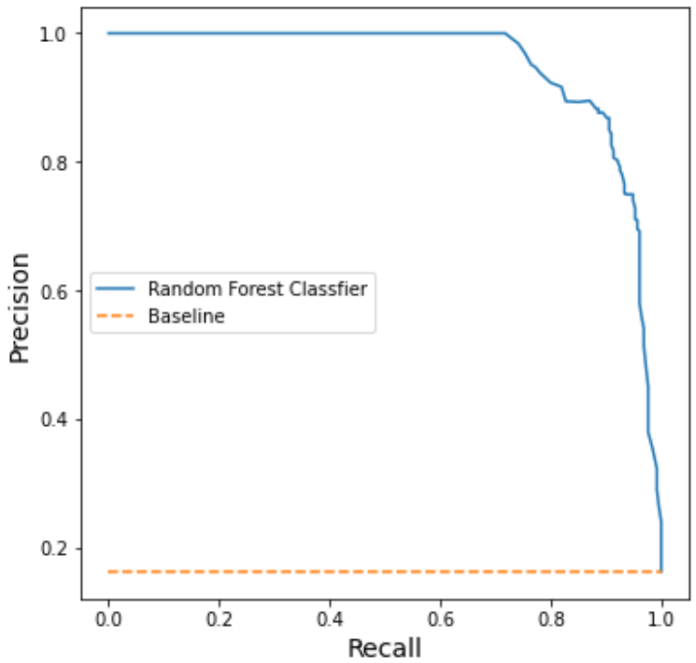


Figure 21. Precision Recall Curve

Figure 21 plots the precision-recall curve of the random forest classifier for predicting whether a loan will fully recover. The AUC is 0.95, which indicates that this classifier has a great performance in this task.

Furthermore, table 4 is the confusion matrix of this classifier when the threshold is 0.9. We can see that the model can perfectly identifies those loans will not fully recover. Therefore, we can expect ensemble will improve our prediction.

	Predict: 0	Predict: 1
Actual: 0	1,318	0
Actual: 1	104	151

Table 4. Confusion Matrix

6.3. Ensemble Model Performance

Finally, we ensemble the random forest classifier and the random forest regressor together and try different thresholds. Figure 22 plots the root mean squared error by the different thresholds of the classifier. We can see that the ensemble model achieves the best RMSE of 0.2217 with a threshold of 0.55. We will use the threshold of 0.55 in the simulation as this is the best hyperparameters we tuned from the training.

In comparison, if we only use the random forest regressor alone to predict the loss, this would achieve an RMSE of 0.2439. We show that the ensemble method can reduce the RMSE by 0.0222, which supports our model design. Furthermore, the 2.22% improvement could really be economically significant when we are working with a large pool of loans.

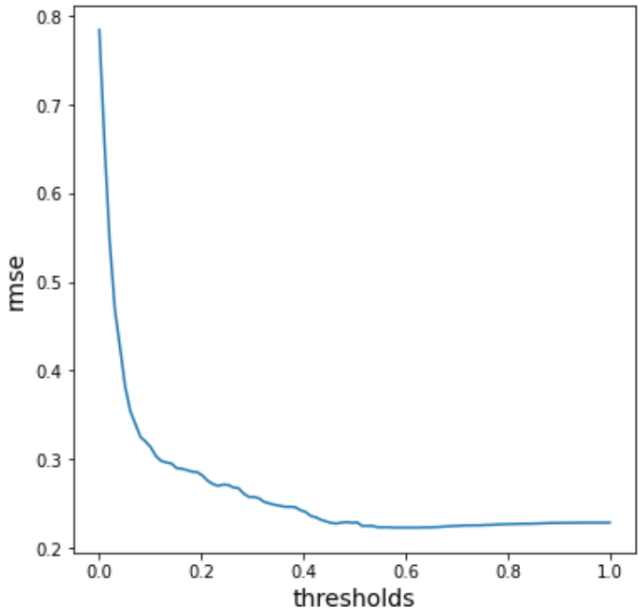


Figure 22. RMSE Curve by Threshold

7. Simulations

In this section, we will simulate the potential risk of our selected portfolio. Section 7.1 introduces how we simulate future macroeconomics. Section 7.2 shows the loss distribution and section 7.3 shows the loss distribution by

different tranches.

7.1. Simulation Methodology

As our time-varying cox model requires quarterly macroeconomic data during the simulation period, we develop a data augmentation method to simulate the data. We choose real macroeconomic statistics during 2005 and 2014, then recursively generate quarterly data by adding a perturbed real data difference to simulated data of last quarter, which is given by

$$Y_{t+1} = Y_t + (1 + r_{long\,term} + r_{long\,term} * r_t) * (X_{t+1} - X_t).$$

Here, $r_{long\,term}$ factor is deterministic for each simulation, describing the long term macroeconomic situation. While r_t factor depicts the short term volatility for each quarter. $\{X_t\}$ are the real macroeconomic data and $\{Y_t\}$ are the simulated data.

7.2. Simulating Loss Distribution

To estimate the loss distribution, we generated simulations of the loan losses for the portfolio in batches. For each batch of 1,450 portfolio simulations, we computed the total losses. Figure 23 shows the total loss distribution for 1,450 portfolio simulations. We can see that the 5-year losses are more concentrated around the central tendency (the mode), thus the 5-year loss is more peaked with a higher degree of kurtosis. On the other hand, the 10-year losses are more spread out in the tails, which means there are more extreme values in the distribution.

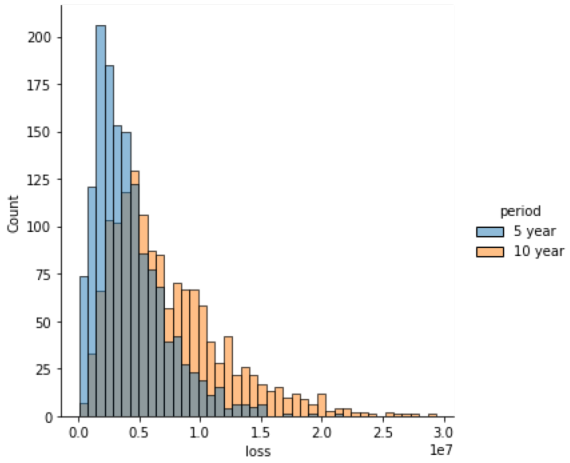


Figure 23. Loss Distribution over 5 and 10 Years

As the loss distribution is clearly not normal, we estimate the empirical Value at Risk at 95% and 99% levels with a 95% confidence interval using bootstrap. We resampled from the total loss with replacement and computed the 95% or 99% quantile loss. We repeated the resampling for 1,000 times and computed the mean and standard deviation of the 95% or 99% quantile losses. Following this procedure, we generated the table 5 that shows the VaR results at the 95% and 99% levels with a 95% confidence interval for 5-years and 10-years, respectively.

		95% Confidence Interval	
	mean	lower	upper
5-year VaR @ 95%	9,907,483	9,377,561	10,437,405
5-year VaR @ 99%	13,779,522	12,867,408	14,691,636
10-year VaR @ 95%	16,753,888	15,925,667	17,582,108
10-year VaR @ 99%	21,802,143	20,371,314	23,232,972

Table 5. Value at Risk

Similarly to our Value-at-Risk estimation procedure, we bootstrapped the same metrics for the Average Value-at-Risk, also called "expected tail loss" (ETL). This metric represents the expected loss on the portfolio in the worst 1% and 5% of scenarios, respectively. We repeated this analysis for 5-year and 10-year simulations. Table 6 summarizes those quantities.

		95% Confidence Interval	
	mean	lower	upper
5-year Avg VaR @ 95%	12,212,593	11,491,142	12,934,045
5-year Avg VaR @ 99%	15,393,033	14,108,448	16,677,618
10-year Avg VaR @ 95%	19,868,832	18,940,411	20,797,253
10-year Avg VaR @ 99%	24,321,599	22,580,422	26,062,775

Table 6. Average Value at Risk

In short, we can find that the total loss over 10 years is larger than that over 5 years and there is more uncertainty over 10 years than 5 years.

7.3. Loss Distribution by Tranche

For the last part of our analysis, we estimated the distribution for the 5-year and 10-year losses of an investor who has purchased a [5%, 15%] tranche backed by the 500-loan portfolio. We also investigated the loss distribution of the [15%, 100%] senior tranche for the given portfolio.

To transform the total loss for each simulation into loss for the tranche, we use the formula

$$f(x) = \begin{cases} 0 & L_{simulation} < a \times L_{total} \\ (b-a) \times L_{total} & L_{simulation} > b \times L_{total} \\ L_{simulation} - a \times L_{total} & otherwise \end{cases}$$

where [a,b] are the bounds of the tranche, L_{total} is the total approval amount of the 500-loan portfolios and $L_{simulation}$ is the total loss in a simulation.

Following this procedure, we compute the loss for the [5%, 15%] tranche. Figure 24 plots the loss distribution for the junior tranche. We can see that the loss over 10 years centered at loss = 1 million; while the loss over

5 years peak at loss $\approx$ 0 million or 1 million. Moreover, the number of simulations that a 10-year portfolio suffers a greater loss is way larger than the number of simulations that a 5-year portfolio suffers loss. From a risk management point of view, a [5%, 15%] tranche backed by the portfolio is a risky investment. When we bought this asset, we may not want to hold it for a longer period as this increases the chances of extreme downturns that lead to a huge loss. Therefore, a risk-averse or risk-neutral investor may not want to hold a junior tranche asset for a long time.

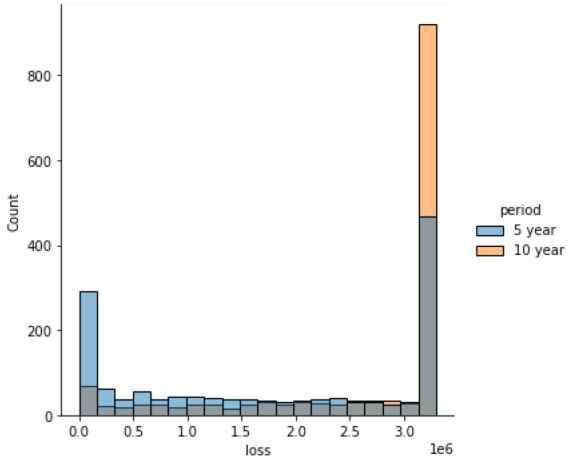


Figure 24. [5%, 15%] Tranche Loss Distribution over 5 and 10 Years

Following a similar procedure, we compute the loss for the [15%, 100%] tranche. Figure 25 plots the loss distribution for the senior tranche. We can see that both of 5-year loss and 10-year loss centered at < 0.2 million. In comparison, the 10-year loss has a higher peak with a longer and fatter tail. A [15%, 100%] tranche backed by the portfolio is a less risky investment. When we bought this asset, we have smaller chances of suffering from huge losses than the junior tranche. However, holding the senior tranche for 10 years is still riskier than holding it for 5 years as there are more uncertainties about the future economics.

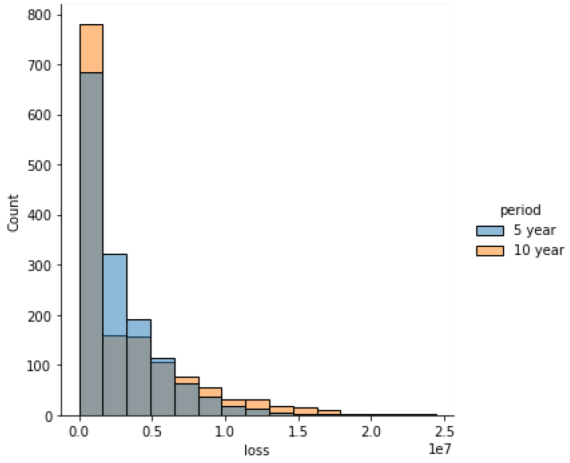


Figure 25. [15%, 100%] Loss Distribution over 5 and 10 Year

In short, a risk-averse would be more willing to purchase a [15%, 100%] senior tranche backed by the portfolio for a shorter period, like 5 years. There are fewer uncertainties about this asset and the investor can expect a smaller chance of incurring a loss with a small amount.

8. Refinement in the Future

Even though our default model works well in general, there are a couple improvements that should be considered in the future.

8.1. Built Additional Model for the Baseline Hazard

Because the baseline hazard is canceled in the partial likelihood, the general economic conditions affecting all loans simultaneously are not considered in the model fitting. One potential way to get around this is to manually estimate the baseline hazard using these country-level macroeconomic data and then manually calculate the estimated cumulative hazards and survivals. In detail, we first compute the observed default rate for each year $\mu(t) = \frac{\#\{X_i\, default\, at\, t\}}{\#\{X_i\, survives\, at\, t\}}$. Under the moderate assumption that the default rate is indifferent with loan ages, the observed default rate converges to the intrinsic baseline default rate for each year. Then, we can fit a generalized additive model (GAM) of observed default rate against the country-level macroeconomic data. With this model, the adjusted accumulated hazard rate has the form:

$$\lambda = \mu(\beta_1; Z_t) \cdot \lambda_{prop}(\beta_2; X_t),$$

i.e. the empirical baseline is replaced with the fitted baseline hazard.

Compared with standard cox model, this adjusted hazard model incorporates systematic risk influenced by the country-level macroeconomic situation, while the standard cox model uses a rather problematic hypothesis that the baseline hazard is only related with the loan age. We believe this adjusted model will greatly improve the prediction accuracy and help to understand the influence of systematic risks in loan-backed asset markets.

8.2. Using Stochastic Differential Equation to Simulate Macroeconomic Scenarios

In our current simulation method, as we want the simulated macroeconomic data to be close to real statistics, the simulation is based on the perturbed real data:

$$Y_{t+1} = Y_t + (1 + r_{long\, term} + r_{long\, term} * r_t) * (X_{t+1} - X_t).$$

Here, $r_{long\, term}$ is deterministic for each simulation, while $r_t \sim \mathcal{N}(0, 0.04)$ is short term fluctuation for each quarter.

However, our current method may restrict the occurrence of extremity events, thus making the loss prediction too optimistic. An alternative method is to use Stochastic Differential Equations to simulate these data, i.e.

$$dY_t = \mu(t)Y_t dt + \sigma(t)Y_t dW_t \tag{2}$$

$$Y(0) = X(0) \tag{3}$$

where $\mu(t)$ and $\sigma(t)$ are the average growth rate and volatility, $X(0)$ is the initial condition, for which we can use the real macroeconomic data in certain year and W_t is a Brownian Motion Process. In algorithm implementation,

we can first randomly generate μ_0 and σ_0, as well as hyperparameter r_μ, r_σ and ρ, then using SDE to simulate a trajectory of $\mu(t)$ and $\sigma(t)$ by

$$\begin{aligned}
\mu(t+\delta t) - \mu(t) &= \mu(t) r_\mu (W^{(1)}_{t+\delta t} - W^{(1)}_t), \\
\sigma(t+\delta t) - \mu(t) &= \mu(t) r_\sigma (W^{(2)}_{t+\delta t} - W^{(2)}_t), \\
dW^{(2)}(t) &= \rho dW^{(1)}(t) + \sqrt{1-\rho^2} dW^{\perp}.
\end{aligned}$$

Then we use (2) to generate macroeconomic data.

9. Conclusion

We explore, in this project, using time-varying cox model to perform default prediction. Our results suggest that the time-varying cox model fits well in this problem setting but is sensible to extreme cases and outlying data. Our model coefficients reveals three key findings with strong economical implications. First, treating the corporate business type as our baseline, the default hazards for personal loans do not differ significantly from that of the baseline business type. This maybe because all business, even labeled as corporation, are small businesses so being a small business does not differ much from being an individual business. Loans for partnership business, on the other hand, have significantly lower default hazard due to their strong relationships with more robust and mature business groups. Second, state and zipcode-level economic factors such as unemployment rate and housing prices are, indeed, significant in predicting defaults. when economic conditions worsens, default hazard increases. Finally, Gross State Product does not have significant implications of whether a specific loan would default or not.

Instead of purely using a regression model to predict the loss, we ensemble a binary classifier and a regressor to predict the loss given default. We exploit the random forest to explore non-linearities and intersections across features. We successfully show that the ensemble model outperforms the regression one.

Finally, we select 500 loans to form a portfolio and try to understand the investment risk behind it. To generate the macroeconomic data for our time-varying cox model, we develop a data augmentation method to simulate these data based on real statistics between 2005 and 2014. Then we use our fitted cox model to simulate default events during the 10 years after origination. The default loans are then passed into our random forest classifier and regressor to predict corresponding losses.

We use the above method to simulate the loss distribution in 5-year and 10-year and compute the Value at Risk and the Average Value at Risk at 95% and 99% levels with a 95% confidence interval using bootstrap. We also estimate the distribution for the 5-year and 10-year losses of the [5%, 15%] junior tranche and the [15%, 100%] tranche backed by the 500-loan portfolio respectively. We find that there are less uncertainties over 5 years and the overall risk for the senior tranche is lower.

References

Kevin W. Bowyer, Nitesh V. Chawla, Lawrence O. Hall, and W. Philip Kegelmeyer. SMOTE: synthetic minority over-sampling technique. *CoRR*, abs/1106.1813, 2011. URL http://arxiv.org/abs/1106.1813.

David Collet. *Model checking in the Cox regression model*. Chapman and Hall/CRC, 3 edition, 2014.

D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972.

Patricia M. Grambsch and Terry M. Therneau. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, 81(3):515–526, 1994.

Patricia M. Grambsch and Terry M. Therneau. *Model survival data: extending the Cox model*. Springer Science + Business Media, LLC, 3 edition, 2000.

Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, 10:e0118432, 03 2015. doi: 10.1371/journal.pone.0118432.

SBA. Loans, 2023. URL https://www.sba.gov/funding-programs/loans.

Terry Therneau, Cynthia Crowson, and Elizabeth Atkinson. Using time dependent covariates and time dependent coefficients in the cox model. 2023.