# MS&E 246 Financial Risk Analytics

**Yifei Guo, Yiting Zhao, Zhengji Yang**

Stanford University

MS&E 246
March 27, 2023

# Outline I



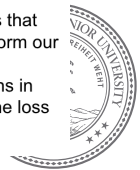
# Outline



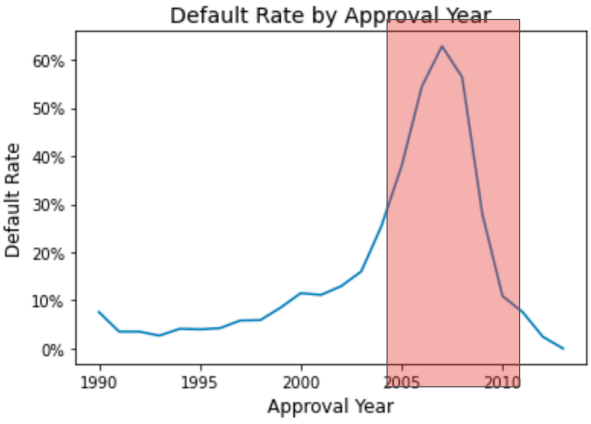
# SBA Dataset Overview

| Feature Name | Feature Engineering |
|---|---|
| Gross Approval | |
| Gross Charge Off Amount | |
| Term In Months | Numerical Variable |
| Project State | Categorical variable with 50 levels. |
| Business Type | Categorical variable with 3 levels |
| Loan Status | Categorical variable with 3 levels |
| Loan Purpose | Take keyword from subprogram; Categorical variable with 4 levels |
| 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. |
| 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. |

- **Data Cleaning**
  - Exclude loans that are canceled or without loan status.
  - Exclude loans with 0 terms.
  - Exclude loans in states other than U.S. 50 mainland states.
- **Data Insights**
  - 8,865 loans default with 1,341 fully recover - **Data Imbalance**
  - 6,039 loans with same end dates - **Ties**
  - 72,014 loans with status: Exempt - **Right Censoring**
- **Portfolio Selection**
  - We randomly select 500 loans that were approved after 2010 to form our 500-loan portfolio.
  - We will not use these 500 loans in fitting the default model and the loss model.

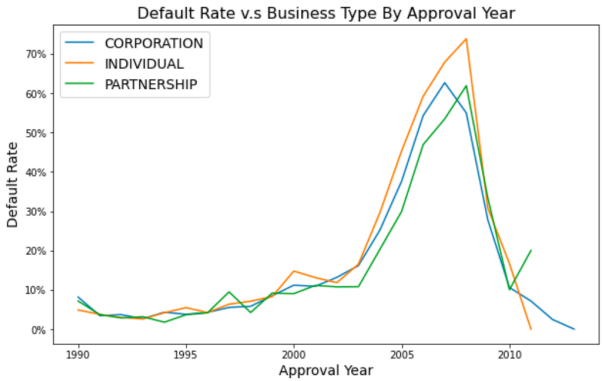


# Default Risk by Approval Year



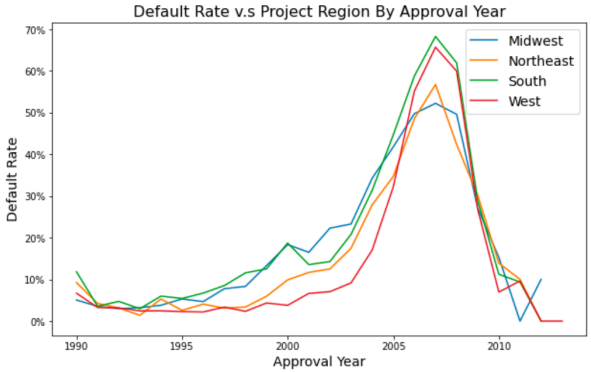

- Default rates spiked for the loans that are approved around the financial crisis (2007-2009)

# Default Risk v.s. Business Type



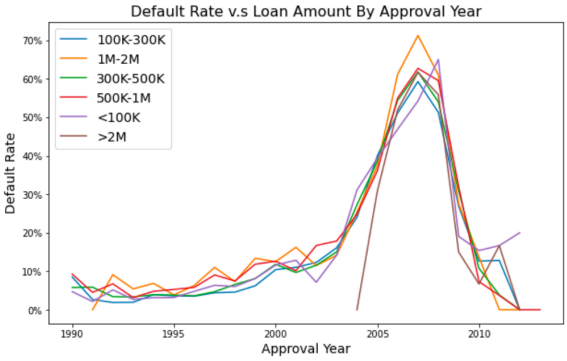

- Different business types were affected differently.
- The “Individual” businesses suffers greater default risk than corporation and partnership.

# Default Risk v.s. Project State




- Projects in different states were affected differently.
- The projects in the South and West suffers greater default risk than those in the Midwest and Northeast.

# Default Risk v.s. Loan Amount




- Loan of different sizes have different default rate patterns over time.
- Loans of sizes $1M - $2M appeared to have experienced larger default risk than other loans in the Great Recession.

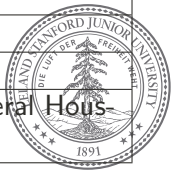
# Outline



# External Data

| 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 |

| 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 |



# Preliminary Model - Time Invariate

```
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
```

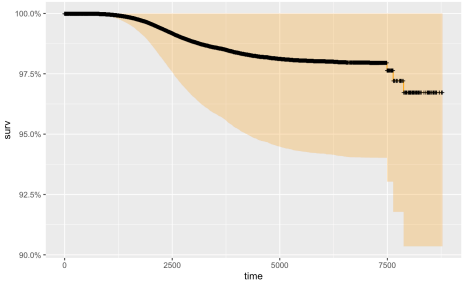



Figure: Results from the preliminary Model

# Preliminary Findings

Findings from this time-invariate proportional hazard model:

- Loan characteristics co-variates are significant: log_amount, TermInMonths, BusinessType, is_Same_Borr, is_Same_Borr_Project
- loan_purpose is not significant (not included in the final model for computational efficiency)
- Country-level macroeconomic data are significant: TED, log_SP, VIX, PRIME, Leverage
- **Assumptions of proportional Hazard?**
- **Linearity?**

# Test Proportional Hazard Hypothesis

The following outcome 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. [1].

| | 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: Results from cox.zph()


Results: A number of the variables and the Global result are significant, indicating that the proportional hazard assumption is violated.

# Preliminary Findings

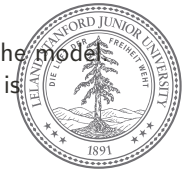
- Loan characteristics co-variates are significant: log_amount, TermInMonths, BusinessType, is_Same_Borr, is_Same_Borr_Project
- loan_purpose is not significant (not included in the final model for computational efficiency)
- Country-level macroeconomic data are significant: TED, log_SP, VIX, PRIME, Leverage
- The proportional hazard assumption does not holde.
  - Remedies? [2]
  - **Stratify**: incorporate co-variates as stratification factors.
  - **Partition the time axis**
  - **Time-dependent co-variates**: $\beta(t)X = \beta X(t)$
  - **Different model**
- **Linearity?**

# Test for Linearity [2]

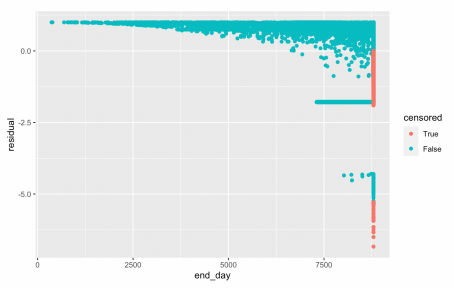
- Estimated a null model with only the intercept term.

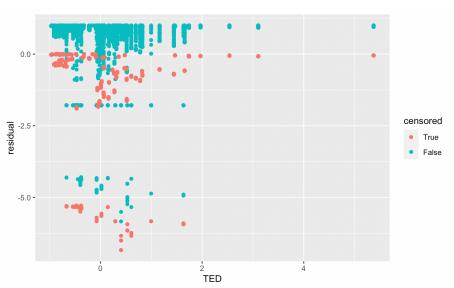
```
Call:  coxph(formula = Surv(time, status == 2) ~ 1, data = dfori)

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

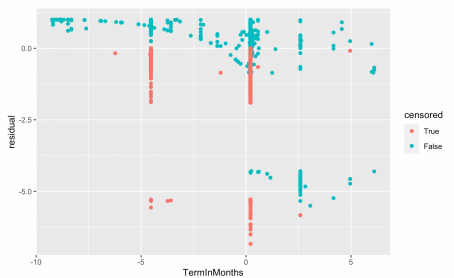
- Calculate martingale residuals of the proportional hazard model.
- Martingale Residual is defined as $MR_i = \delta_i - r_{ci}$, where $r_{ci}$ is based on the Nelson Aalen estimate of the cumulative hazard function (pp.115). [3]
- Properties of Martingale Residual:
  - $\mathbf{E}(MR_i) = 0$, $\sum \hat{M}_i = 0$, $cov(M_i, M_j) = 0$, $cov(\hat{M}_i, \hat{M}_j) < 0$ [2].
  - 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.
- Plot the martingale residuals against the co-variates.

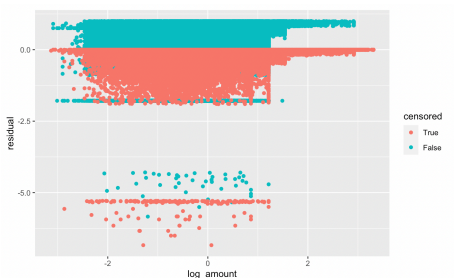


# Plots of Martingale Residuals

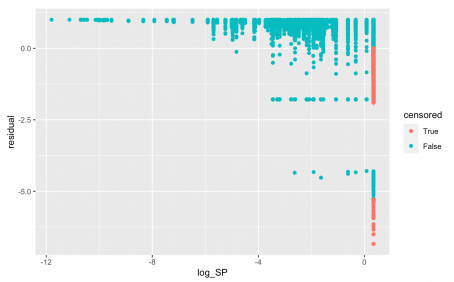


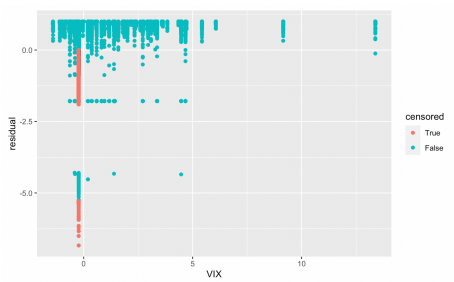



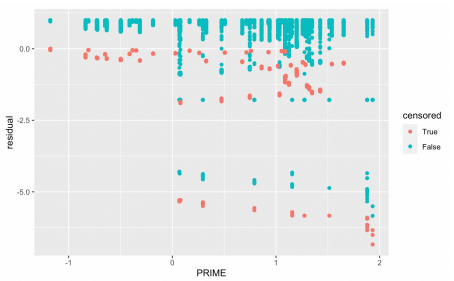



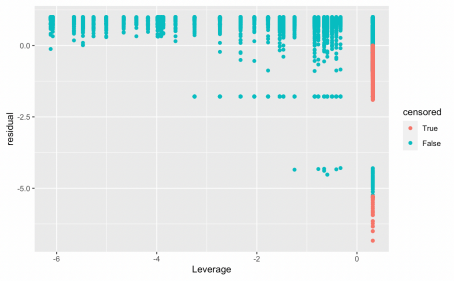



# Plots of Martingale Residuals



# Interpretation: Plots of Martingale Residuals

- 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.
- 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.

# Preliminary Findings

- Loan characteristics co-variates are significant: log_amount, TermInMonths, BusinessType, is_Same_Borr, is_Same_Borr_Project
- loan_purpose is not significant (not included in the final model for computational efficiency)
- Country-level macroeconomic data are significant: TED, log_SP, VIX, PRIME, Leverage
- The proportional hazard assumption does not hode.
  - What to do [2]?
  - **Stratify**: incorporate co-variates as stratification factors.
  - **Partition the time axis**
  - **Time-dependent co-variates**: $\beta(t)X = \beta X(t)$
  - **Different model**
- Need to add nonlinear term for TermInMonths.
- Also need to consider right-censoring and ties

# Preliminary Findings

- Include loan features except loan_purpose
- Add state and zipcode level economic data as time varying co-variates.
- Add nonlinear term for TermInMonths.
- Need to consider right-censoring and ties.

# Outline



# Recap: Cox Model without Censoring [2]

- Let $T_i^*$ denote the time from loan origination to default.
- **Assumption 1**: $T_i^*$ are iid with pdf $f(t)$,.
- Let $S_t = P(T^* > t)$ denotes the survival function, then the hazard function is $\lambda(t) = lim_{h \to 0} P(t \leq T^* < t + h \mid T^* \geq t)/h = \frac{f(t)}{S(t)}$
- Cox Model with time varying co-variates: $\lambda_i(t) = \lambda_0(t) e^{X_i(t)\beta}$.
- Hazard ratio: $\frac{\lambda_i(t)}{\lambda_j(t)} = \frac{\lambda_0(t) e^{X_i(t)\beta}}{\lambda_0(t) e^{X_j(t)\beta}} = \frac{e^{X_i(t)\beta}}{e^{X_j(t)\beta}}$
- Let $Y_i(t)$ be an indicator of whether unit i is under observation and at risk at time t.
- Because the proportional hazard assumption is not satisfied by the preliminary model, we incorporate time-varying co-variates into our model.

# Define the model with right censoring [2]

- In our data set, 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.
- **Assumption 2 (non-informative assumption)**: $C_i^*$ is independent of $T_i^*$.
- Then the likelihood with censoring is:

$$\prod_{i:\, default} f(T_i^*|X_i) \times \prod_{j:\, PIF} F(T_j^*|X_j) \times \prod_{k:\, censor} F(C^*|X_k)$$

$$= \prod_{i:\, default} \lambda(T_i|X_i) \times \prod_{j=1}^{n} F(T_j|X_j)$$

where $F(t|X_j) = F(\tau \leq t - t_j|X_j; \beta) = exp(-\int_{t=t_j}^{t} \lambda(X_j, \beta)\, ds)$.

# Tied Data [2]

There are a total of 6039 loans with tied end dates. Because there are ties within our dataset, we cannot assume that each event time corresponds to exactly one event.

- **Why this matter?**
- According to the to Cox [4], the estimation of $\beta$ is based on the partial likelihood function: $PL(\beta) = \prod_{i=1}^{n} \prod_{t \geq 0} \frac{Y_i(t) exp[X_i(t)\beta]}{\sum_j Y_j(t) exp[X_j(t)\beta]}$
- **What can we do?**
- Breslow approximation: the least accurate approximation.
- **Efron approximation**: better and more complicated than Breslow while having simple computation.
- Exact partial likelihood: exhaustive enumeration but changes the model's functional form to a logistic model: $\frac{\lambda_i(t)}{1+\lambda_i(t)} = \frac{\lambda_0(t)}{1+\lambda_0(t)} exp(X_i(t)\beta)$
- Averaged likelihood: exhaustive enumeration with numerical evaluation of integrals.

# Efron approximation [2]

To compute the partial likelihood for tied data, 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.

# Nonlinearity [2]

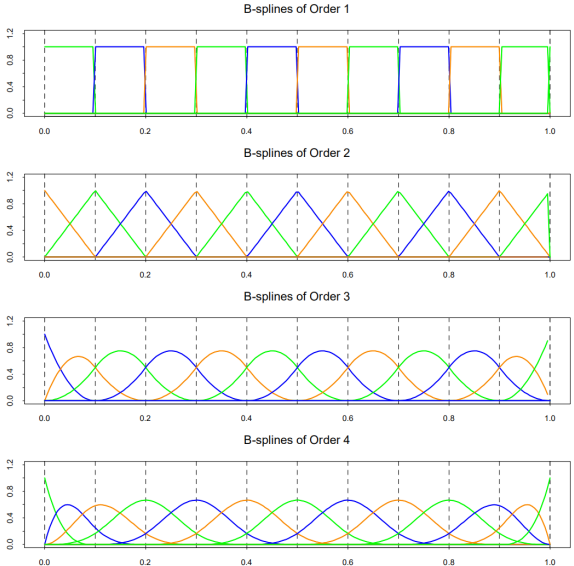
To model the nonlinear effect of some features on the hazard rate, we use Penalized Smoothing Splines (psplines) for prediction. The smoothing spline (of some order) is

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

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.

- Flexible: By choosing proper N, $f(\cdot)$ can be a good approximation to any function.
- Robust: $B_i(\cdot)$ has local support, hence $f(\cdot)$ will not diverge to infinity beyond the sample range.

# B-spline Basis [2]



# pspline and penalty [2]

As the spline can be quite complex and may overfit, a roughness penalty $J(f)$ is added into the objective 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 $\beta$ 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 $f_j$.

# 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 [5]. 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. cluster(id) is added to the model code to specify which loan does a row correspond to.

# Snapshot of our Time-Varying Dataset

| 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: First 20 lines of out dataset

# Fitting Algorithm

- The survival package implemented the Newton-Raphson Algorithm to solve the partial likelihood equation.
- Starting with an initial guess $\hat{\beta}^0$
- Compute iteratively until convergence: $\hat{\beta^{n+1}} = \hat{\beta}^n + \mathcal{I}^{-1}(\hat{\beta}^n)\, U(\hat{\beta}^n)$

# Outline



# Default Model

- 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 propotional hazard rate.

- The loss function with penalty is:

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

# Default Model Fitting

```
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: Results from our default model

Recap likelihood-based test:

- The likelihood ratio test statistics: $2 * l(\hat{\beta} - \hat{\beta}_0)$
- The Wald test statistics: $(\hat{\beta} - \hat{\beta}_0)^T \hat{\mathcal{I}} (\hat{\beta} - \hat{\beta}_0)$, where $\hat{\mathcal{I}}$ is the information matrix.

# Baseline Hazard

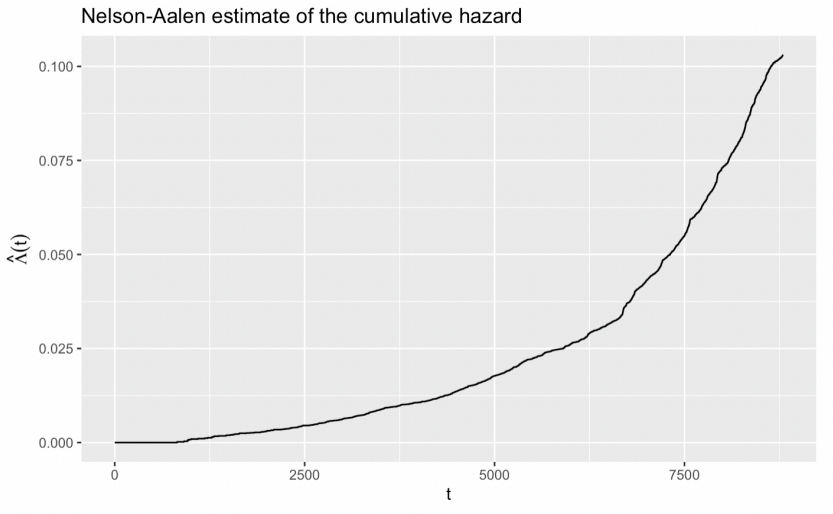



Figure: Nelson-Aalen estimate of the cumulative hazard

# Survival Curve

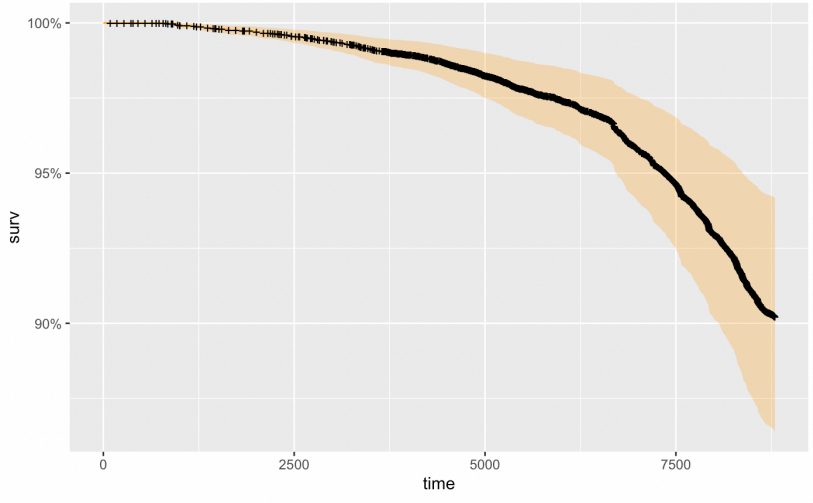
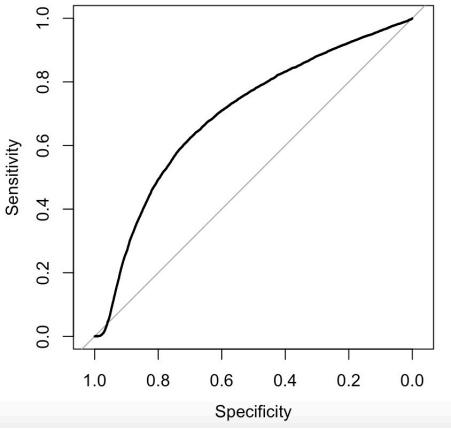



Figure: Estimated Survival Curve

# Training Dataset vs. Testing Dataset



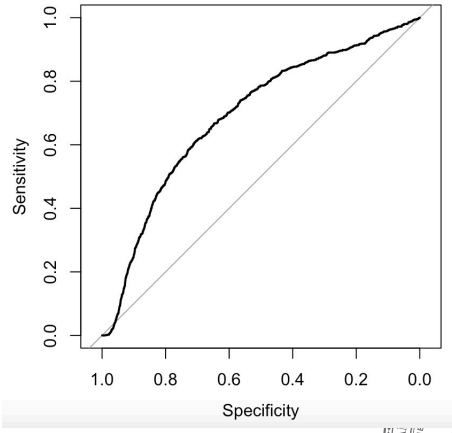
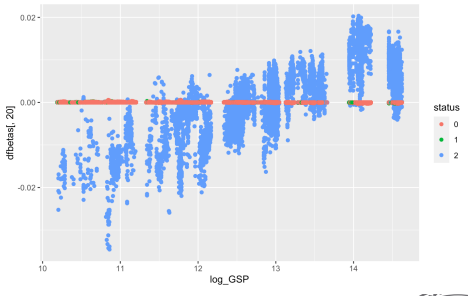




Figure: ROC Curve for training dataset and for testing dataset (AUC for Training: 0.726; AUC for Testing: 0.724)

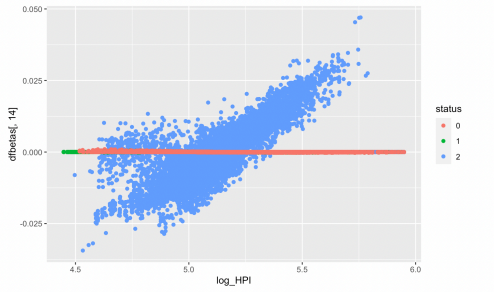
# Influences

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.


- 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.

# Influences




- 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.

# Outline



# Train-Test Split

- We only use 8,865 defaulted loans, after excluding those in the 500-loans portfolio.
- In addition to loan characteristics, we also use additional data incl. SP500, VIX, TED, PRIME, Leverage, GSP, PersonalIncome, UnemploymentRate, IndustryGDP, 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 on the remaining data.

# Model Design

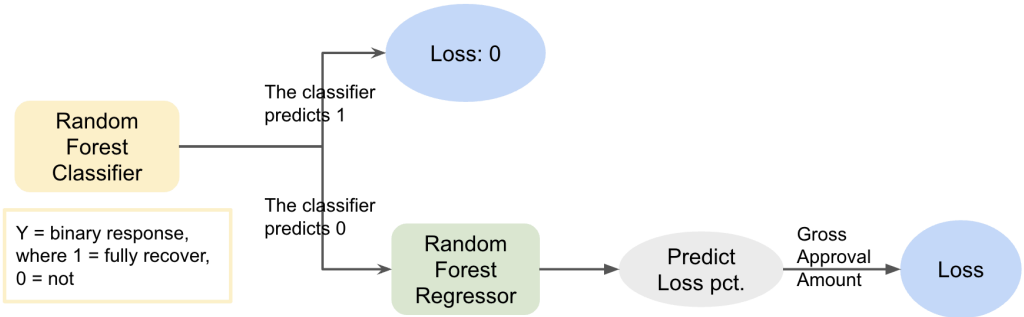
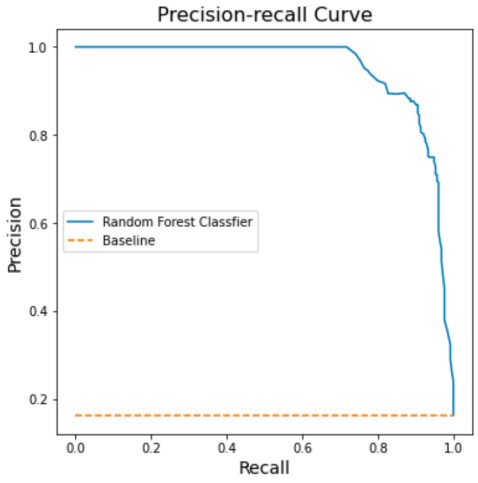



Figure: Loss Model Design

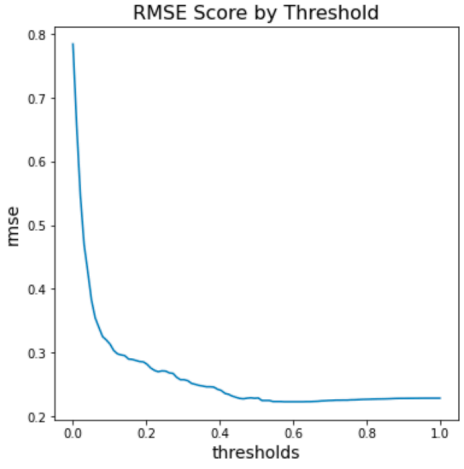
# Evaluation: Classifier




- Data Imbalance: only ~15% of the loans in the train set are fully recovered
  - Oversampling minority class with SMOTE and random undersampling the majority class
  - Use Precision Recall Curve to evaluate the model performance
- AUC = 0.95
- Confusion Matrix (threshold = 0.9)

| | Predict 0 | Predict 1 |
|---|---|---|
| Actual 0 | 1318 | 0 |
| Actual 1 | 104 | 151 |

# Evaluation: Combined Results




- The ensemble model achieves the best RMSE of 0.2217 with threshold of 0.53
- While the regresor alone can achieve RMSE of 0.2439
- **The ensemble method can further reduce the RMSE by 0.0222**

# Outline



# Simulation

For the analysis of pool level risk, we randomly choose 500 loans originated after 2010 and simulate the default events in ten years after origination. We use our cox model to generate default event, i.e. whether default and when, then feed the data into the loss model to predict loss for each default loan.

Producing time varying features

- loan related data, e.g. TermInMonths, GrossApprovalAmount, NAICS Code,... are fixed
- time varying features are generated based on the real data from 2005-2014
- we randomly generate a condition factor $r_{fixed} \in (-1, 1)$, which describes whether the economic situation is relatively good in the whole period.
- we generate data for each quarter recursively using $X_{i+1} = X_i + (R_{i+1} - R_i) * (1 + r_{fixed} * \tau_X + r_{random})$.

# Simulation

Simulating default events

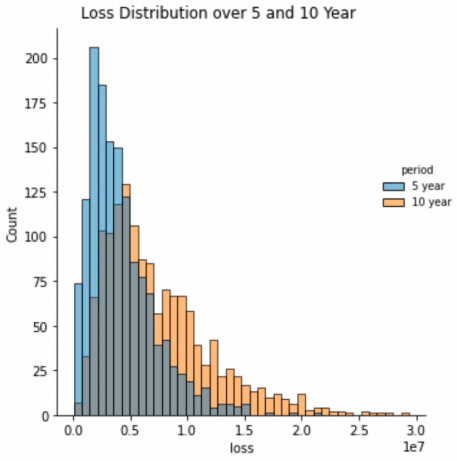
- Using the fitted cox model to predict proportional hazard $\lambda_{prop}(X_i, t; \beta)$ for each loan in each quarter
- Computing the survival function $S(t) = exp(-\int_{s=0}^{t} \lambda_0(s)\lambda_(X_i, t; \beta)\, dt)$ for each quarter.
- Generate uniform r.v. $Unif(0, 1)$, then determining default using $U > S(T_0)$, $T_0$ is 5yr or 10yr.
- Collect all default loan for each simulation and record default times $\tau = min\{t : U > S(t)\}$.

# Simulation

Computing pool level loss

- Passing the data into random forest classifier to determine whether a default loan is fully recovered.
- Collecting partially recovered loans and predict loss
- Compute gross amount of loss for each simulation.

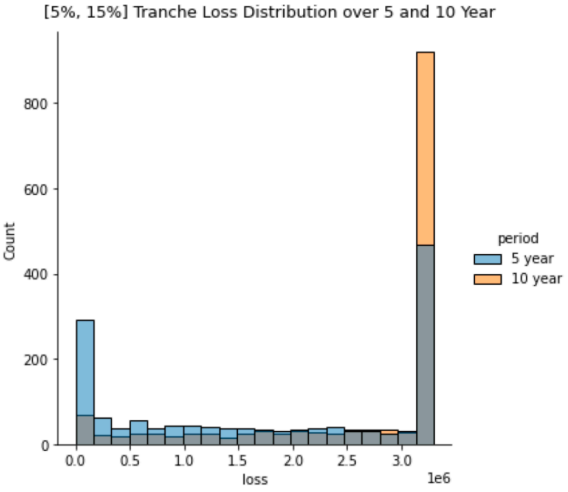
# Loss Distribution: 5-year v.s 10-year




| | | 95% Confidence Interval | |
|---|---|---|---|
| | mean | lower | upper |
| 5 yr VaR @ 95% | 9,907,483 | 9,377,561 | 10,437,405 |
| 5 yr VaR @ 99% | 13,779,522 | 12,867,408 | 14,691,636 |
| 10 yr VaR @ 95% | 16,753,888 | 15,925,667 | 17,582,108 |
| 10 yr VaR @ 99% | 21,802,143 | 20,371,314 | 23,232,972 |
| 5 yr Avg VaR @ 95% | 12,212,593 | 11,491,142 | 12,934,045 |
| 5 yr Avg VaR @ 99% | 15,393,033 | 14,108,448 | 16,677,618 |
| 10 yr Avg VaR @ 95% | 19,868,832 | 18,940,411 | 20,797,253 |
| 10 yr Avg VaR @ 99% | 24,321,599 | 22,580,422 | 26,062,775 |

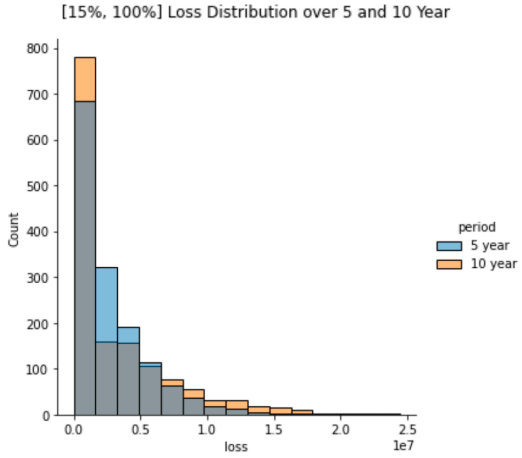
- The total loss over 10 year is larger than that over 5 year.
- There is more uncertainty over 10 year than 5 year.

# Loss Distribition ([5%, 15%] tranche): 5-year v.s 10-year




- The loss over 10 year centered at loss = 1 million; while the loss over 5 year peaked at loss ~0 million or = 1 million
- 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 to avoid a larger loss.

# Loss Distribition ([15%,100%] tranche): 5-year v.s. 10-year




- Both of 5-year loss and 10-year loss centered at < 0.2 million; 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. 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.

# Bibliography I


[1] P. M. Grambsch **and** T. M. Therneau, "Proportional hazards tests and diagnostics based on weighted residuals," *Biometrika*, **jourvol** 81, **number** 3, **pages** 515–526, 1994.

[2] P. M. Grambsch **and** T. M. Therneau, *Model survival data: extending the Cox model*, 3 **edition**. Springer Science + Business Media, LLC, 2000.

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

[4] D. R. Cox, "Regression models and life-tables," *Journal of the Royal Statistical Society. Series B (Methodological)*, **jourvol** 34, **number** 2, **pages** 187–220, 19[illegible]

[5] T. Therneau, C. Crowson **and** E. Atkinson, "Using time dependent covariates and time dependent coefficients in the cox model,", 2023.


# Thank you!


LELAND STANFORD JUNIOR UNIVERSITY
DIE LUFT DER FREIHEIT WEHT
1891