Modelling Dynamic Prepayment and Default with Survival Analysis and Machine Learning in Credit Protection Insurance

Alessia Eletti, Marco Aleandri
Sapienza Università di Roma

May 24th 2019
Agenda

- Background
- Market dynamics and actuarial model
- Survival analysis with machine learning
- Dynamic lapse rate and TVOG
- Study case: CPI on US loans
Agenda

- Background
- Market dynamics and actuarial model
- Survival analysis with machine learning
- Dynamic lapse rate and TVOG
- Study case: CPI on US loans
Background

- Richardson et al., Lapse rate.
- Outreville, Whole-life insurance lapse rates and the emergency fund hypothesis.
- Loisel et al., From deterministic to stochastic surrender risk models: impact of correlation crises on economic capital.
- Barsotti et al., Lapse risk in life insurance: correlation and contagion effects among policyholder behaviors.
- Nolte et al., Don’t lapse into temptation: a behavioral explanation for policyholder surrender.
# Background

**Emergency Fund Hypothesis**

Lapse risk is mainly driven by a natural – and irrational to some extent – response to the need of money due to personal conditions in time of distress

**Interest Rate Hypothesis**

Lapse risk is mainly driven by rational reasonings (e.g., interest rate arbitrage and preference for different products) due to the policyholder's risk appetite

<table>
<thead>
<tr>
<th>Policyholder-related factors</th>
<th>Product-related factors</th>
<th>Macroeconomic factors</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Age</td>
<td>- Duration</td>
<td>- Market yields</td>
</tr>
<tr>
<td>- Family</td>
<td>- Premium</td>
<td>- Unemployment rate</td>
</tr>
<tr>
<td>- Salary</td>
<td>- Guaranteed rate</td>
<td>- Gross domestic product</td>
</tr>
<tr>
<td>- ...</td>
<td>- ...</td>
<td>- ...</td>
</tr>
</tbody>
</table>
Agenda

- Background
- Market dynamics and actuarial model
- Survival analysis with machine learning
- Dynamic lapse rate and TVOG
- Study case: CPI on US loans
Market dynamics and actuarial model

\[ dr_t = a(b - r_t)dt + \sigma dZ_t \]

\[ v(t, T) = -\frac{1}{T-t}lnA(t, T) + \frac{1}{T-t}B(t, T)r(t) \]

\[ g_t = \alpha + \beta r_t \]

- Vasicek model is chosen so as to allow for negative interest rates
- Parameters are considered as additional surrender drivers

\[ I_t = (r_{t-1} + s)\frac{(1+i)^t}{(1+i)^n-1}C_0 \]

\[ P = C_0U_{x:n\bar{\mu}} = C_0\frac{IA_{x:n\bar{\mu}}}{\bar{a}_{x:n\bar{\mu}}} \]

\[ V_t = C_tU_{x+t:n-t\bar{\mu}} \]

Note: the mathematical reserve is deterministic as the stochastic interest rate doesn't affect the capital amortization
Agenda

- Background
- Market dynamics and actuarial model
- Survival analysis with machine learning
- Dynamic lapse rate and TVOG
- Study case: CPI on US loans
Survival analysis with Machine Learning

**Accelerated Failure Time model**

\[ \log(T) = \beta'x + \sigma \epsilon \]

\[ S(t \mid x) = S_0(t \cdot \exp(-\beta'x)) \]

\[ h(t \mid x) = h_0(t \cdot \exp(-\beta'x)) \exp(-\beta'x) \]
### Survival analysis with Machine Learning

**Accelerated Failure Time model**

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Survival function</th>
<th>Hazard function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weibull</td>
<td>$S(t) = \exp(-\lambda t^p)$</td>
<td>$h(t) = \lambda p t^{p-1}$</td>
</tr>
<tr>
<td>Log-logistic</td>
<td>$S(t) = \frac{1}{1 + \exp(\theta)t^\kappa}$</td>
<td>$h(t) = \frac{\exp(\theta)\kappa t^{\kappa-1}}{1 + \exp(\theta)t^\kappa}$</td>
</tr>
<tr>
<td>Log-normal</td>
<td>$S(t) = 1 - \Phi\left(\frac{\ln x - \mu}{\sigma}\right)$</td>
<td>$h(t) = \frac{1}{x\sigma} \Phi\left(\frac{\ln x - \mu}{\sigma}\right)$</td>
</tr>
<tr>
<td>Gamma</td>
<td>$S(t) = 1 - \frac{\gamma(\alpha, \beta x)}{\Gamma(\alpha)}$</td>
<td>$h(t) = \frac{x^{\alpha-1}e^{-\beta x}}{\Gamma(\alpha) - \gamma(\alpha, \beta x)}$</td>
</tr>
</tbody>
</table>
Survival analysis with Machine Learning

Survival Random Forest

B bootstrap samples are drawn from the original data.
Survival analysis with Machine Learning

Survival Random Forest

\[ \text{node } h \]

\[ \text{node } j = 1 \]
\[ p \]

log balance orig
log interest rate means
gdp beta
log FICO
uer orig
investor orig
...

\[ \text{node } j = 2 \]
Survival Random Forest

\[ x^* \geq c^* \quad \text{node} \quad j = 1 \]

log-rank test

\[ |L(x^*, c^*)| \geq |L(x, c)| \]

\[ x^* < c^* \quad \text{node} \quad j = 2 \]
Survival analysis with Machine Learning

Survival Random Forest

log-rank statistic

\[ L(x, c) = \frac{\sum_{i=1}^{N} \left( d_{i,1} - Y_{i,1} \frac{d_i}{Y_i} \right)}{\sqrt{\sum_{i=1}^{N} \frac{Y_{i,1}}{Y_i} \left( 1 - \frac{Y_{i,1}}{Y_i} \right) \left( \frac{Y_i - d_i}{Y_i - 1} \right) d_i}} \]

ordered unique death times

\[ t_1 < t_2 < \ldots < t_N \]

number of deaths at \( (j, t_i) \)

\[ d_{i,j} \]

number of individuals at \( (j, t_i) \)

\[ Y_{i,j} \]

\[ \hat{\Lambda}_j(t) = \sum_{t_{i,j} < t} \frac{d_{i,j}}{Y_{i,j}} \]
Survival Random Forest

\[
\hat{\Lambda}_j^{(E)}(t) = \frac{1}{B} \sum_{b=1}^{B} \hat{\Lambda}_j^{(b)}(t)
\]
Agenda

- Background
- Market dynamics and actuarial model
- Survival analysis with machine learning
- Dynamic lapse rate and TVOG
- Study case: CPI on US loans
Dynamic lapse rate and TVOG

Surrender probability

\[ \hat{Q} \equiv \frac{\hat{S}(t) - \hat{S}(t + 1)}{\hat{S}(t)} \]

Lapse rate

\[ \hat{l}_r := 1 - \left[ 1 - \hat{Q}_P(r) \left( 1 - \frac{\hat{Q}_D(r)}{2} \right) \right] \left[ 1 - \hat{Q}_D(r) \left( 1 - \frac{\hat{Q}_P(r)}{2} \right) \right] \]

Certainty equivalent profit

\[ D_{ce} = f(i, q_x; r_{ce}, C_0, s, \beta) \]

Stochastic profit

\[ D_j = f(i, q_x; r_j, C_0, s, \beta) \]

Average stochastic profit

\[ \overline{D}_j = \frac{1}{N} \sum_{j=1}^{N} D_j \]
\[ D_{ce} > \bar{D}_j \]
Agenda

- Background
- Market dynamics and actuarial model
- Survival analysis with machine learning
- Dynamic lapse rate and TVOG
- Study case: CPI on US loans
Study case: CPI on US loans

Dataset and data pre-processing

• Contains origination and performance information on 50,000 borrowers synthesised in cross-section form

• Left and right censored data with maximum loan observation time of 60 trimesters, i.e. 15 years

• Features
  • origination balance (logarithm)
  • interest rate origination value, mean and variance (logarithm)
  • unemployment rate and gdp origination value
  • unemployment rate and gdp period-wise indicator
  • logarithm of FICO credit score
  • real estate and investor type
Study case: CPI on US loans

Parametric vs non-parametric model

• Model 1: Accelerated Failure Time model with Weibull distribution.

• Model 2: Random Survival Forest.

• Logistic regression.

• When evaluating the two models, censored data is dealt with by computing an AUC for each time \( t \).
Study case: CPI on US loans

Parametric vs non-parametric model
Study case: CPI on US loans

Parametric vs non-parametric model

Default prediction (train)

- Weibull AFT
- random forest
- logistic regression

AUC
Study case: CPI on US loans

Parametric vs non-parametric model

Payoff prediction (train)

AUC

trimester

- Weibull AFT
- random forest
- logistic regression
Study case: CPI on US loans

Parametric vs non-parametric model

\[ AUC = \frac{\sum_{t=1}^{N} AUC_t n_t}{\sum_{t=1}^{N} n_t} \]

<table>
<thead>
<tr>
<th></th>
<th>Default</th>
<th>Prepayment</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Train</td>
<td>Validation</td>
</tr>
<tr>
<td>AFT model</td>
<td>71%</td>
<td>73%</td>
</tr>
<tr>
<td>Random Forest</td>
<td>94%</td>
<td>91%</td>
</tr>
<tr>
<td>Logistic</td>
<td>71%</td>
<td>64%</td>
</tr>
<tr>
<td>regression</td>
<td>71%</td>
<td>64%</td>
</tr>
</tbody>
</table>

AFT model vs Random Forest vs Logistic regression
Study case: CPI on US loans

Default and prepayment estimation
Study case: CPI on US loans

Default and prepayment estimation
Study case: CPI on US loans

TVOG analysis

Fig. 32: TVOG shape by $a$.  
Fig. 33: TVOG shape by $\sigma$.  
Fig. 34: TVOG shape by $b$.  
Fig. 35: TVOG shape by $r(0)$.  
Fig. 36: TVOG shape by $s$.  
Fig. 37: TVOG shape by $\beta$.  
Study case: CPI on US loans

Limitations, extensions and conclusion

• Random forest **outperforms** the AFT model (*through-the-cycle model*). This is due to the fact that RF relies on a higher number of degrees of freedom.

• Random forest also **outperforms** the logistic regression (*point-in-time model*). Indeed, in the latter stepwise selection of features at each time leads to further loss of information.

• The **significantly better performance** of the random forest comes at the cost of **very slight overfitting**. This is typical as ensemble learners tend to fit both linear and more complex non linear relations while keeping low generalization error.

• TVOG tends to vary between 0% and 6% for all parametrizations seen, except for values of $\beta$ higher than the baseline value. Negative interest rates and the dual nature of default, in fact, can lead to negative TVOG.
Study case: CPI on US loans

Limitations, extensions and conclusion

• **Cox model**: the main goal of this study was to confront traditional parametric model and semi-parametric random forest but we could have chosen Cox proportional hazard model (which is also a common survival model) as the counterpart in the confrontation with RF.

• **Competing risks model**: our model doesn’t take into consideration the correlation between default and prepayment. This is partly justified by the fact that these options tend to be the consequence of opposite situations for the policyholder. CRM account for their interaction and thus are a possible extension of our model.

• **Mixture cure modelling**: in credit-risk modelling a large part of the population never defaults/prepays, hence the survival probability will never actually be zero but will level at some value.
Any questions?