

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

BSc Research Dissertation / Scientific Paper

Dept. of Statistical Sciences

Univ. “La Sapienza”, Rome

Alessia Eletti, BSc Candidate
Email: alessiaeletti@gmail.com

Dr. Marco Aleandri, Project Tutor
Email: marco.aleandri@uniroma1.it

Prof. Fabio Grasso, Project Supervisor
Email: fabio.grasso@uniroma1.it

May 6, 2019

1 Introduction

The objective of this paper is to describe and predict policyholder behaviour dynamically to evaluate its impact on the profit of a credit protection insurance product (CPI) in the US market. In particular, we will assess how macroeconomic factors and contractual characteristics influence the duration and outcome of the loan taken out by the individual, i.e. whether it ends by default or prepayment. The estimations will be performed by traditional regression-based survival analysis as well as tree-based machine learning techniques, in order to highlight benefits and drawbacks of both.

Broadly speaking, CPI is aimed to protect the lender in the eventuality that the borrower dies or faces severe disability preventing him from servicing the debt. As such, the first step of the analysis involves building a stochastic profit model for that. It will encompass a stochastic interest rate model to simulate the actual macroeconomic environment, the actuarial model for the CPI and the financial model for the underlying variable-rate loan.

While some elements are deterministic assumptions (e.g., mortality, loan value, etc.) or predefined deterministic functions of the stochastic interest rate (e.g., loan amortization, fund rate, etc.), prepayment and default rates will be estimated starting from the dataset used in Baesens et al. (2016). A natural approach relates to survival analysis, but it will be implemented in two different ways:

- Weibull regression
- Survival random forest

assuming the parameters of the underlying distributions as target variables. Each of them will predict separated survival functions for prepayment and default, which will evolve as the interest rate scenario changes, contributing to the dynamic nature of the profit model.

Finally, we will assess the impact of prepayment and default on the profit. As they can be seen as put options embedded within the CPI, the profit will be necessarily affected by the time value of options and guarantees (TVOG), providing us with a marginal measure of the option value.

2 Background

Lapse risk has always been considered of great importance to an insurance company's health. Richardson et al. (1951) add that it's important for the agent and the policyholder as well. High early terminations, in fact, frequently result in heavy loss to the company and are one of the major causes of dissatisfaction among policyholder with the life insurance business. Since then much attention has been devolved to the estimation of lapse risk. Many studies have been conducted in order to determine the factors which mainly influence the chance of surrender. Among others Outreville (1990) tested the hypothesis that cash values are utilized by policyholders as an emergency fund, certainly not a new idea in insurance literature yet scarcely tested at the time, as opposed to the interest rate hypothesis, still predominant to date. The reasons at the base of lapse behavior are not considered only a byproduct of external or internal

45 economic factors but are also traced back to individual decision heuristics, fi-
46 nancial literacy and financial advice, as explored by Nolte et al. (2017). The
47 search for more complex and well rounded explanations for lapsing can be linked
48 to the recent economic crises. On the same line, contagion phenomena are also
49 starting to gain interest. Barsotti et al. (2016) leverage on a specific extension
50 of Hawkes process, the so-called dynamic contagion process, to capture both
51 contagion and correlations potentially arising among policyholders' behaviors.
52 Note that in this framework the lapse intensity process depends once again on
53 the interest rate trajectory, which changes as a consequence of a market driven
54 jump. Finally, the characteristics of the new Solvency II framework once again
55 stress once again the importance of lapsing and, thus, of its correct estimation.
56 The necessity to determine the surrender chance is in fact relevant both in life
57 and non-life, as lapse represents a sub-module which enters in the calculation
58 of the SCR for both. A better estimation of lapsing thus leads to a more realis-
59 tic estimation of the capital requirement with respect to the company's actual
60 needs. For all of these reasons, numerous models have been discussed and used
61 through the decades. Further, as this paper explores mortgage default and pre-
62 payment in connection with a particular form of life related insurance as stated
63 in the introduction, the analysis conducted by Li (2014) with regards to loan
64 related lapse is noteworthy. It in fact draws upon academic literature on resi-
65 dential mortgage default and research on stress testing published by regulatory
66 bodies to identify four statistical models that can be used to estimate chance
67 of surrender for and individual loan. These are linear regression analysis on de-
68 fault risk, logistic models, survival analysis and optimization models. Note that
69 the logistic model is one of the preferred paths for Baesens et al. (2016). Par-
70 ticularly, in a discrete-time hazard model setting, logistic regression is used to
71 estimate 0-1 policyholder behavior by linking it FICO credit score, loan-to-value
72 and gross domestic product. In continuous-time hazard model setting, instead,
73 a survival model is chosen, notably an accelerated failure time model with ex-
74 ponential survival, a simpler version of the model used in this paper also with
75 respect to the variables chosen. With regards to the optimization model, this
76 simply consists in the assumption that the borrower makes mortgage payment
77 decisions with the objective of maximizing wealth and utility and minimizing
78 housing related costs. At any given point of time the possible choices for the
79 borrower thus include continuing with the current mortgage, prepaying or de-
80 faulting and various wealth effects are associated with each of the choices, as
81 explored by Capozza et al. (1996).

82 It's clear that lapse rate linkage to interest rate trajectories has been abun-
83 dantly emphasized throughout decades of literature and practical application.
84 Further, following on the idea of stress testing, a final element vital to a com-
85 plete lapse rate estimation is that of linking it to various interest rate scenarios
86 and studying it's evolution accordingly. Stress testing has in fact acquired great
87 importance both in the banking world and in Solvency II. As a consequence of
88 the establishment of ORSA as part of the company's risk-management system,
89 it's required to assess not only current but also prospective risk. Notably as can
90 be found in the Level 3 consultation paper, Guideline 9, (paragraph 3.24) on
91 the assessment of overall solvency needs:

92 *The undertaking should subject the identified risks to a sufficiently*
93 *wide range of stress test/scenario analyses to provide an adequate*

94

basis for the assessment of the overall solvency needs.

95 In this perspective, it's clear how a stochastic take on modelling interest rates
 96 and thus chance of surrender can be necessary or at least more insightful com-
 97 pared to a deterministic approach. Notably, Loisel et al. (2010) raise the matter
 98 of considering a stochastic model for the surrender rate instead of the classi-
 99 cal S-shaped deterministic curve, based on the spread between the interest rate
 100 given by the contract and the one that the policyholder could obtain somewhere
 101 else in the market. This paper, though, then concentrates on quantifying the
 102 impact of the correlation phenomenon on a real life portfolio under a partially
 103 internal Solvency II model. The present paper, on the other hand, exploits the
 104 simulation of numerous interest rate scenarios through a Vasicek stochastic pro-
 105 cess to dynamically determine lapse rate and consequentially premium to profit
 106 proportion and TVOG. Further, the estimation is done combining the afore-
 107 mentioned survival model with an ensemble machine learning approach thus
 108 combining multiple of the current interests revolving around chance of surren-
 109 der and mentioned in this background research.

110 3 Market dynamics and actuarial model

111 With regards to the simulation of the risk-free interest rates, a traditional Va-
 112 sicek one-factor model is chosen in order to allow for negative interest rates.
 113 Indicating with r the short rate, the model has the following form:

$$dr_t = a(b - r_t)dt + \sigma dZ_t \quad (1)$$

114 which is characterized by mean-reverting drift and standard deviation of the
 115 change in the short term proportional to the standard Brownian motion dZ_t
 116 (see Hull (2012)). Rather than calibrating it on historical market rights, we will
 117 consider its parameters as additional drivers of defaults and prepayments. This
 118 is reasonable as interest rate levels represent their main trigger.

119 The simulated curves play a number of roles in the profit model. They are used
 120 to calculate risk-free discounting factors, performance of the assets backing the
 121 technical provisions (defined by a linear function of the risk-free interest rate)
 122 and base forward rates for the loan installments (calculated as the sum of the
 123 risk-free interest rate and a contractual fixed spread). More specifically, given
 124 the amortization rate i , the contractual spread s and the initial loan capital C_0 ,
 125 the installment of the month t is calculated as follows:

$$I_t = (r_{t-1} + s)C_{t-1} = (r_{t-1} + s) \frac{(1+i)^t}{(1+i)^n - 1} C_0 \quad (2)$$

126 that is the sum of a deterministic principal component and a stochastic interest
 127 component.

128 As the CPI is typically provisioned by a single initial payment, the single pre-
 129 mium and the reserve amount in t are calculated as follows:

$$P = C_0 U_{\overline{x:\overline{n}|i}} = C_0 \frac{IA_{\overline{x:\overline{n}|i}}}{\ddot{a}_{\overline{x:\overline{n}|i}}}, \quad V_t = C_t U_{\overline{x+t:n-t|i}} = C_t \frac{IA_{\overline{x+t:n-t|i}}}{\ddot{a}_{\overline{x+t:n-t|i}}} \quad (3)$$

130 using the traditional actuarial notation. Notice that the mathematical reserve
 131 is deterministic as the stochastic rate does not affect the capital amortization.

132 4 Survival analysis with machine learning

133 If $T \sim Exp(\lambda)$ is the random variable representing the duration of the loan,
 134 the marginal survival function (i.e., the probability that at a given month t the
 135 individual has not terminated his loan, either by default or by prepaying) is
 136 defined by $S(t) = Pr(T > t)$. As already mentioned, prepayment and default
 137 rates are estimated separately by two different survival functions defined by
 138 different λ s. For sake of simplicity, however, we will neglect such a distinction
 139 below.

140 Three different methods will be used: exponential regression, survival tree and
 141 survival forest. While the former is a traditional, parametric method, the other
 142 two are alternative, non-parametric techniques often used in machine learning
 143 and data science.

144 With regards to the regression approach, a generalized linear model based on
 145 the exponential distribution is being considered for the logarithmic survival time
 146 of the k^{th} policyholder:

$$\ln T_k = -\mathbf{x}'_k \boldsymbol{\beta} + \epsilon_k \quad (4)$$

147 so that the estimated hazard function for that record can be written as follows:

$$\hat{\lambda}_k = \hat{\lambda}_0 e^{-\mathbf{x}'_k \hat{\boldsymbol{\beta}}}. \quad (5)$$

148 When considering a survival tree, instead, a one-step full likelihood method
 149 for proportional hazard models (see M. LeBlanc et al. (1992)) is being used
 150 as implemented in the R package `rpart` (see T. Therneau et al. (2010)). In
 151 particular, expressing the assumption of proportional hazards model as $\lambda_h(t) =$
 152 $\theta_h \lambda_0(t)$, the likelihood of the given tree is

$$L = \prod_{h \in \tilde{T}} \prod_{k \in S_h} [\theta_h \lambda_0(t)]^{\delta_k} e^{-\Lambda_0(t_k) \theta_h} \quad (6)$$

153 where h indicates the node, k indicates the individual, $\delta_k := I_{T_k}$, \tilde{T} is the set of
 154 terminal nodes and S_h is the set of observation labels in the h^{th} terminal node.
 155 Given the baseline hazard function $\Lambda_0(\cdot)$, the maximum likelihood estimates of
 156 $\{\theta_h : h \in \tilde{T}\}$ are

$$\tilde{\theta}_h = \frac{\sum_{k \in S_h} \delta_k}{\sum_{k \in S_h} \Lambda_0(t_k)} \quad (7)$$

157 where generally $\Lambda_0(\cdot)$ is not known, but can be estimated as result of an alter-
 158 nating estimation procedure. In particular, at every j^{th} iteration, the following
 159 quantity can be defined

$$\hat{\Lambda}_0^j(t) = \frac{\sum_{k:t_k \leq t} \delta_k}{\sum_{h \in \tilde{T}} \sum_{k:t_k \geq t, k \in S_h} \hat{\theta}_h^j}. \quad (8)$$

160 Only the first iteration will be used in the recursive partitioning procedure to
 161 grow and select the tree size. All in all, the one step estimate of θ_h is thus given
 162 by

$$\hat{\theta}_h^1 = \frac{\sum_{k \in S_h} \delta_k}{\sum_{k \in S_h} \hat{\Lambda}_0^1(t_k)} \quad (9)$$

163 and the hazard rate $\hat{\lambda}_h(t)$ can be determined recursively step by step.

164 Although survival trees tend to outperform regression models, their non-parametric

165 estimations are often affected by significant instability. As a consequence, tree-
 166 based ensembles (i.e., combinations of a number of different trees grown up
 167 from the same dataset) such as random forests are generally preferred in order
 168 to mitigate this problem.

169 Survival forests are considered here as defined in H. Ishwaran et al. (2007) and
 170 H. Ishwaran et al. (2008). This approach is based on the idea that all aspects
 171 of growing forest must take into account the outcome. In right censored survival
 172 settings such as this one, it means that survival time and censoring informa-
 173 tion have to be explicitly included in the splitting criterion used. Further, the
 174 predicted value for a terminal node in a tree, the measure of the effectiveness
 175 of a split and the measure of prediction accuracy must all properly incorporate
 176 survival information.

177 As a first step, the algorithm draws B bootstrap samples from the original data.
 178 Then a survival tree is grown for each bootstrap sample. In particular, at each
 179 node p covariates are randomly selected and used in the splitting process. This
 180 is done so as to maximize the survival difference between daughter nodes. Note
 181 that the introduction of these two sources of randomization contribute to the
 182 further reduction of the generalization error. The default way to measure such
 183 difference in the R package `ranger` is to use the log-rank test defined as

$$L(x, c) = \frac{\sum_{i=1}^N \left(d_{i,1} - Y_{i,1} \frac{d_i}{Y_i} \right)}{\sqrt{\sum_{i=1}^N \frac{Y_{i,1}}{Y_i} \left(1 - \frac{Y_{i,1}}{Y_i} \right) \left(\frac{Y_i - d_i}{Y_i - 1} \right) d_i}} \quad (10)$$

184 for a split at value c for the predictor x . The measure of node separation is given
 185 by $|L(x, c)|$. The best split at node h is determined by finding the predictor x^*
 186 and the split value c^* such that, for each x and c ,

$$|L(x^*, c^*)| \geq |L(x, c)|. \quad (11)$$

187 The final steps of the algorithm consist of growing each tree to full size under the
 188 constraint that a terminal node should have no less than a predefined number
 189 of unique events. The hazard function is then calculated for each tree. More
 190 specifically, let $d_{l,h}$ be the number of prepayments or defaults and $Y_{l,h}$ the
 191 number of active individuals at time $t_{l,h}$, where l indicated the number of unique
 192 event times. The hazard function estimate for h is given by

$$\hat{\Lambda}_h(t) = \sum_{t_{l,h} \leq t} \frac{d_{l,h}}{Y_{l,h}} \quad (12)$$

193 which is the hazard function shared by all the policyholders belonging to h and
 194 returned by a single survival tree. To compute the bootstrap ensemble hazard
 195 function, we need to average over the B hazard functions from the B survival
 196 trees:

$$\hat{\Lambda}_h^{(E)}(t) = \frac{1}{B} \sum_{b=1}^B \hat{\Lambda}_h^{(b)}(t). \quad (13)$$

197 5 Dynamic lapse rate, profit and TVOG

198 Comparing the three aforementioned estimations with common classification
 199 measures such as the ROC curve or the cross entropy on training and validation

200 data, we will select the model representing the best trade-off between accuracy
 201 and stability. The estimation itself is, in turn, preparatory to the determination
 202 of the probability that the policyholder defaults or prepays in a certain month
 203 t , given that until t he remained in the portfolio. In other words, given that
 204 the model returns the hazard function and thus the survival function $S(\cdot)$, we
 205 obtain the following conditional probabilities of prepayment and default:

$$\widehat{Q}_P \equiv \frac{\widehat{S}_P(t) - \widehat{S}_P(t+1)}{\widehat{S}_P(t)}, \quad \widehat{Q}_D \equiv \frac{\widehat{S}_D(t) - \widehat{S}_D(t+1)}{\widehat{S}_D(t)} \quad (14)$$

206 which are constant for each t due to the choice of the exponential distribution,
 207 but vary by policyholder and interest rate scenario. All in all, the overall lapse
 208 rate is given by

$$\widehat{l}_r := 1 - \left[1 - \widehat{Q}_P(r) \left(1 - \frac{\widehat{Q}_D(r)}{2} \right) \right] \left[1 - \widehat{Q}_D(r) \left(1 - \frac{\widehat{Q}_P(r)}{2} \right) \right] \quad (15)$$

209 where r formally recalls the dependence to the short rate and thus the stochastic
 210 model. Given that the actual risk-free curve represents the certainty equivalent
 211 scenario, it will be used to obtain the related certainty equivalent lapse rate. At
 212 the same time, a given stochastic simulation j will lead to a specific lapse rate.
 213 In formulae:

$$\widehat{l}_{ce} = 1 - \left[1 - \widehat{Q}_P(r_{ce}) \left(1 - \frac{\widehat{Q}_D(r_{ce})}{2} \right) \right] \left[1 - \widehat{Q}_D(r_{ce}) \left(1 - \frac{\widehat{Q}_P(r_{ce})}{2} \right) \right] \quad (16)$$

$$\widehat{l}_j = 1 - \left[1 - \widehat{Q}_P(r_j) \left(1 - \frac{\widehat{Q}_D(r_j)}{2} \right) \right] \left[1 - \widehat{Q}_D(r_j) \left(1 - \frac{\widehat{Q}_P(r_j)}{2} \right) \right] \quad (17)$$

214 for each j . In parallel, we will distinguish three different definition for the dis-
 215 counted profit of the CPI: the certainty equivalent profit $D_{ce, \widehat{l}_{ce}}$, the stochastic
 216 profit with certainty equivalent lapse rate $D_{j, \widehat{l}_{ce}}$ and the full stochastic profit
 217 D_{j, \widehat{l}_j} . The related average discounted profits are

$$\overline{D}_{\widehat{l}_{ce}} = \frac{1}{N} \sum_{j=1}^N D_{j, \widehat{l}_{ce}} \quad (18)$$

$$\overline{D}_{\widehat{l}_j} = \frac{1}{N} \sum_{j=1}^N D_{j, \widehat{l}_j} \quad (19)$$

218 over N simulations. Although no explicit financial option (e.g., minimum guar-
 219 anteed rate) is embedded in the CPI, prepayment and default represent put
 220 options linked to the policyholder behaviour. This means that the following
 221 relations should hold:

$$D_{ce, \widehat{l}_{ce}} \approx \overline{D}_{\widehat{l}_{ce}} > \overline{D}_{\widehat{l}_j} \quad (20)$$

222 so that the difference between $\overline{D}_{\widehat{l}_{ce}}$ and $\overline{D}_{\widehat{l}_j}$ is a measure of the combined value
 223 of those options, that is, the aforementioned TVOG.

224 **6 Study case: Credit Protection Insurance on** 225 **US loans**

226 This section describes the path that has been followed to analyze how profit
227 and TVOG vary as its predictors vary. As mentioned in the introduction, we
228 will apply two survival models, a Weibull regression and a random forest, to the
229 dataset present in Baesens et al. (2016). The best performing model will be
230 chosen to estimate the lapse rate of a given policyholder profile. This will then
231 be used in a CPI model, together with other elements, in order to determine
232 relevant quantities such as profit, premium/profit ratio and TVOG. Further, we
233 will draw conclusions on the impacts of the predictors on profitability by varying
234 its influencing factors or, equivalently, by varying the policyholder's profile.

235 **6.1 Dataset and data preprocessing**

236 The reference dataset was originally in panel form, and contained origination
237 and performance information on 50.000 borrowers, censored on the right as well
238 as on the left. The maximum loan observation time is equal to 60 trimesters,
239 that is, fifteen years. They were taken out between a minimum origination time
240 of -40 to a maximum maturity time of 60. This corresponds to the period that
241 goes from January 1990 to January 2015. Further, the dataset is characterized
242 by 622.489 observations of 23 variables. Here's a brief overview:

- 243 • `id`: borrower ID
- 244 • `time`: time stamp of observation
- 245 • `orig_time`: time stamp for origination
- 246 • `first_time`: time stamp for first observation
- 247 • `mat_time`: time stamp for maturity
- 248 • `balance_time`: outstanding balance at observation time
- 249 • `LTV_time`: loan-to-value ratio at observation time, in %
- 250 • `interest_rate_time`: interest rate at observation time, in %
- 251 • `hpi_time`: house price index at observation time, base year = 100
- 252 • `gdp_time`: gross domestic product growth at observation time, in %
- 253 • `uer_time`: unemployment rate at observation time, in %
- 254 • `REtype_CO_orig_time`: real estate type condominium = 1, otherwise = 0
- 255 • `REtype_PU_orig_time`: real estate type planned urban development = 1,
256 otherwise = 0
- 257 • `REtype_SF_orig_time`: single-family home = 1, otherwise = 0
- 258 • `investor_orig_time`: investor borrower = 1, otherwise = 0
- 259 • `balance_orig_time`: outstanding balance at origination time

- 260 • `FICO_orig_time`: FICO score at origination time (reference scoring is: <
- 261 620 is bad, 620-649 is poor, 650-699 is fair, 700-749 is good, \geq 750 is
- 262 excellent)
- 263 • `LTV_orig_time`: loan-to-value ratio at origination time, in %
- 264 • `Interest_Rate_orig_time`: interest rate at origination time, in %
- 265 • `hpi_orig_time`: house price index at origination time, base year = 100
- 266 • `default_time`: default observation at observation time (0/1)
- 267 • `payoff_time`: payoff observation at observation time (0/1)
- 268 • `status_time`: default (1), payoff (2), and non default/non payoff (0) at
- 269 observation time

270 Through a preliminary analysis of the dataset we take care of missing values
 271 and rescale or truncate certain variables to render them homogeneous in terms
 272 of their order of magnitude and to reduce skewness and kurtosis. In particular,
 273 there are 270 missing values, all in `LTV_time`, this problem is solved by sub-
 274 stituting them with their predicted values from the regression of `LTV_time` on
 275 `LTV_orig_time`. Furthermore, this variable is characterized by strong skewness
 276 due to only 71 observations with loan-to-value ratio above than 200%, these are,
 277 thus, eliminated. The other transformations done to take care of skewness and
 278 kurtosis involve the variables `interest_rate_time` and `balance_time` and their
 279 counterparts at origination time `interest_rate_orig_time` and `balance_orig_time`,
 280 which are rescaled using a logarithmic transformation. The same is done to
 281 `FICO_orig_time` which is otherwise characterized by a different order of mag-
 282 nitude compared to the other variables. The principal synthetic measures are
 283 reported in ??.

284 The data has then been synthesized into cross-section form by maintaining only
 285 significant information and building indicators that summarize the evolution of
 286 macroeconomic factors during the loan amortization period. In particular, the
 287 indicator is given by the coefficient of the regression of each time-dependant
 288 economic variable (`hpi_time`, `uer_time`, `gdp_time` and `LTV_time`) over `time`. In
 289 this way we obtained the variables `hpi_beta`, `uer_beta`, `gdp_beta` and `LTV_beta`.
 290 Note that we include loan-to-value among the macroeconomic indicators since
 291 the value of the house for which the loan is taken out depends on the overall
 292 economic situation. Non surprisingly we will find it to be highly correlated with
 293 `hpi_time`. In addition to these trend indicators we maintain also their origi-
 294 nation values (`hpi_origs`, `uer_origs`, `gdp_beta` and `LTV_origs`), which are the
 295 intercepts of the regressions. Finally, we also build a variable which represents
 296 the duration of the observation period, defined as the difference between the last
 297 and first observation times plus one and indicated with `remaining_time`. Note
 298 that the survival models will be fitted on this time measure and not on `time`.
 299 The reference dataset from here forward will always be this synthesized version
 300 of the original one. Further, in light of what we explained so far, part of this
 301 information (that related to interest rates) is stochastic. Before concluding this
 302 section and moving on to the description of the models, we analyze the most
 303 interesting correlations (see fig. 1). This will give a preliminary idea of which
 304 features will be selected for the models. The four macroeconomic indicators are

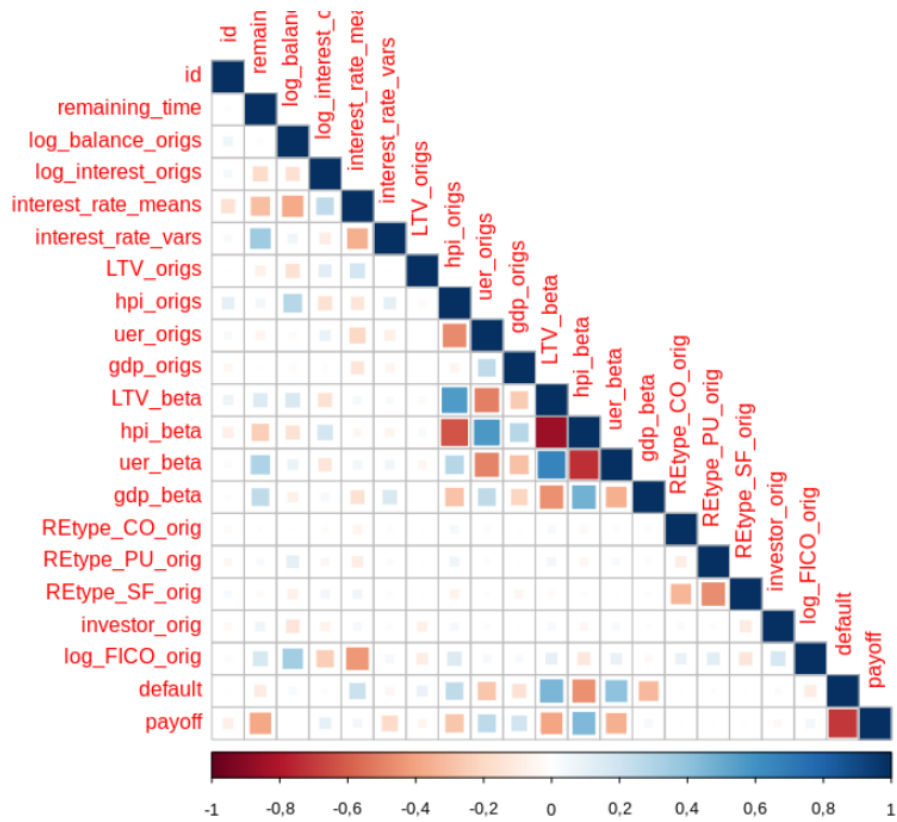


Fig. 1: Correlation matrix (among the time variables, only `remaining_time` was included as it is the most significant one)

305 characterized by relevant reciprocal correlations. Non surprisingly, `LTV_beta`
 306 is positively correlated with `uer_beta` and negatively correlated with `gdp_beta`
 307 and `hpi_beta`. In turn, `uer_beta` is negatively correlated with `gdp_beta` and
 308 `hpi_beta` is negatively correlated with `uer_beta` and positively with `gdp_beta`.
 309 All of these correlations are simply confirmations of the relation we typically see
 310 between these indicators. Other noteworthy correlations include those between
 311 `default` and the four macroeconomic time-wise indicators. Non surprisingly
 312 it's negatively correlated with the house price index trend (`hpi_beta`), and with
 313 the the gross domestic product trend (`gdp_beta`). Further, decreasing loan-
 314 to-value trend (`LTV_beta`) means on average lower default rates, the same is
 315 true for the unemployment rate (`uer_beta`). The opposite is true for the pre-
 316 payment option. Increasing `LTV_beta` and `uer_beta` and decreasing `hpi_beta`
 317 bring on average a lower chance of prepayment. Further, the logarithm of the
 318 FICO credit score is positively correlated with the logarithm of the initial loan
 319 balance. This means that on average policyholders with a higher, i.e. better,
 320 credit score also take out higher loans and vice versa. Instead, it's negatively
 321 correlated with the `interest_rate_means`, i.e. with the average logarithm of the
 322 interest rate paid by the policyholder during the duration of the mortgage.
 323 In turn `interest_rate_means` is positively correlated with the the logarithm

324 of the interest rate origination value (`log_interest_origs`) and negatively cor-
325 related with both `log_balance_origs` and the `remaining_time`. Finally, note
326 that default and prepayment rates are highly and negatively correlated, though
327 this will not be taken into consideration in the model and represents one of its
328 limitations.

329 At this point, the path chosen to analyze policyholder behaviour is that of sur-
330 vival analysis based on the Weibull distribution. In this way, in fact, it will be
331 possible to determine, at each month of the fifteen-year time period, the overall
332 probability of exiting the loan for each policyholder.

333 6.2 Model confrontation: parametric vs non-parametric

334 Once the dataset has been preprocessed, the variables that will enter in the
335 final models are selected. This is done by building a survival object, separately
336 for default and prepayment, and by regressing it on all the remaining variables
337 of the dataset and the additional macroeconomic indicators. Stepwise selection
338 and considerations on VIF (done so as to take care of multicollinearity) are then
339 applied to the regression, thus resulting in the following variables for a given
340 individual:

- 341 • `log_balance_origs`: logarithm of origination balance
- 342 • `interest_rate_means`: logarithm of average interest rate
- 343 • `interest_rate_vars`: logarithm of interest rate variance
- 344 • `log_interest_origs`: logarithm of origination value
- 345 • `uer_origs`: unemployment rate origination value
- 346 • `gdp_origs`: gross domestic product origination value
- 347 • `uer_beta`: unemployment rate period-wise indicator
- 348 • `gdp_beta`: gross domestic product period-wise indicator

349 and the previously described `REtype_CO_orig`, `REtype_PU_orig`, `REtype_SF_orig`,
350 `investor_orig` and `log_FICO_orig`.

351 Note that some popularly used predictors when trying to estimate lapse rate in
352 loans, i.e. loan-to-value ratio and house price index, are excluded from the final
353 model. The reason for this is that they are highly correlated with the other two
354 macroeconomic variables, unemployment rate and gross domestic product, as
355 observed in Section 6.1. This means that as soon as `uer_beta` and `gdp_beta` are
356 selected by the model, `LTV_beta` and `hpi_beta` will lose most of their predictive
357 power, leading to their rejection. The same is true for the corresponding origi-
358 nation variables, `LTV_origs` and `hpi_origs`.

359 We can now look into the survival regression in order to gain information on
360 how the features influence the survival curves, both with regards to default as
361 to prepayment (see table 1 and table 2). In particular, when it comes to the
362 default model, all of the variables are significant at over 99% confidence level
363 except for `gdp_origs` which is significant at 90% confidence level. The Weibull
364 shape (p) and scale (l) parameters, as defined in the R function `rweibull`, are

	Value	Std. error	z	p
(Intercept)	1.455	0.120	12.12	0.000
log_balance_origs	-0.086	0.003	-29.62	0.000
log_interest_origs	-0.010	0.002	-5.18	0.000
interest_rate_means	-0.497	0.010	-49.94	0.000
interest_rate_vars	0.338	0.041	8.18	0.000
uer_origs	7.632	0.460	16.61	0.000
gdp_origs	-0.542	0.285	-1.90	0.057
uer_beta	40.058	1.760	22.76	0.000
gdp_beta	93.419	1.045	89.43	0.000
REtype_CO_orig	-0.062	0.009	-7.23	0.000
REtype_PU_orig	-0.062	0.007	-8.79	0.000
REtype_SF_orig	-0.030	0.005	-5.85	0.000
investor_orig	-0.028	0.006	-4.67	0.000
log_FICO_orig	0.305	0.020	15.17	0.000
Log(scale)	-1.457	0.006	-228.71	0.000

Table 1: Survival regression summary for defaults

	Value	Std. error	z	p
(Intercept)	6.216	0.132	47.22	0.000
log_balance_origs	-0.048	0.003	-15.54	0.000
log_interest_origs	-0.052	0.002	-28.39	0.000
interest_rate_means	-0.368	0.010	-36.40	0.000
interest_rate_vars	2.840	0.083	34.04	0.000
uer_origs	2.768	0.340	8.13	0.000
gdp_origs	-7.593	0.283	-26.86	0.000
uer_beta	236.350	2.224	106.27	0.000
gdp_beta	87.065	1.507	57.79	0.000
REtype_CO_orig	-0.058	0.009	-6.14	0.000
REtype_PU_orig	-0.056	0.008	-7.27	0.000
REtype_SF_orig	-0.037	0.005	-6.83	0.000
investor_orig	0.049	0.007	7.08	0.000
log_FICO_orig	-0.434	0.022	-19.63	0.000
Log(scale)	-1.096	0.005	-215.14	0.000

Table 2: Survival regression summary for prepayments

	Frequency	AFT	Forest	Logistic
default	2.44%	1.80%	2.15%	2.44% (?)
payoff	4.27%	2.43%	3.02%	4.27% (?)

Table 3: Confrontation between surrender frequencies and their estimated probabilities as per the models described above.

365 $p = \frac{1}{scale} = 4.293$ and $l = e^{-linearpredictors}$. The latter varies for each pol-
 366 icyholder which means that, as expected, we have a different survival curve
 367 for every individual. Note also that the coefficients are those of a regression
 368 with logarithmic survival time over negative predictors, as described in Section
 369 4. This means that the effects have to be read by changing the sign of the
 370 coefficients. In this way we obtain results coherent to our expectations. With
 371 regards to the random forest, instead, the parameters have been chosen through
 372 trial and error. The number of trees is set to 100, minimum node size is set
 373 to 100, maximum tree depth is set to 5 and the number of variables randomly
 374 chosen at each split is set equal to 2. Note that the default value of the last
 375 parameter is the rounded down square root of the number of variables. In this
 376 case this would be 3. Note also that the model is fitted on the variables as the
 377 survival regression. Once we've obtained the survival curves for each individual
 378 by fitting the two survival models as described above we evaluate them and see
 379 which one performs better. In particular the conditional probabilities (14) de-
 380 fined in Section 5, are determined for each individual by using the fitted model
 381 on a series of partial versions of the complete dataset. At each trimester t only
 382 policyholders which have begun and continue to be observed at that time are
 383 considered for prediction, i.e. `first_time` $\geq t$ and `default_time` $\leq t$. The same
 384 is done when analyzing `payoff_time`. In other terms, the individuals present
 385 in aggregated form in the cross-section dataset are now taken in their original
 386 representation. The estimations are thus done on a time-wise ordered version of
 387 the panel dataset. On a side note, we also consider a classic logistic regression
 388 on particular partial versions of the original panel dataset. Namely, every group
 389 of statuses characterized by the same observation time are considered a stand-
 390 alone datasets and logistic regression is carried out on each of them both for
 391 default and prepayment on all of the remaining variables. A particular model is
 392 then selected through step-wise selection and the conditional surrender proba-
 393 bilities are computed. The confrontation between the average estimated default
 394 and prepayment probabilities and the default and prepayment frequencies which
 395 derive from the reference dataset is rather noteworthy as a preliminary evalua-
 396 tion of the three models (see table 3). The conditional surrender probabilities
 397 deriving from the survival forest are closer to the actual surrender proportions
 398 than are those estimated through the AFT model. Note that the main con-
 399 frontation will always be between the survival models. The logistic regression,
 400 although interesting as a reference point, is only of marginal importance.
 401 A more complete evaluation of the models is carried out by computing and
 402 confronting the areas under the curve of the respective ROCs. The same sub-
 403 datasets used to estimate the conditional probabilities as stated above are now
 404 used to compute the AUC. In particular, at each trimester t the AUC is deter-
 405 mined on the predictions corresponding to the t^{th} sub-dataset. In this way we
 406 hope to eliminate information redundancy, typical with time-dependant AUCs,

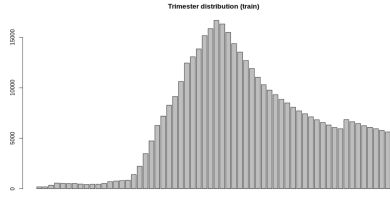


Fig. 2: Trimester distribution, i.e. number of observations in the t^{th} sub-dataset (training)

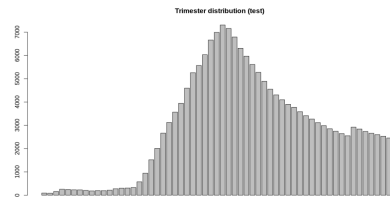


Fig. 3: Trimester distribution, i.e. number of observations in the t^{th} sub-dataset (validation)

407 as pointed out in Kamarudin et al. (2017) with respect to cumulative/dynamic
 408 time-dependant ROCs. The statuses composing each sub-dataset are in fact
 409 mutually exclusive and globally amount to the panel version of the dataset.
 410 Note also that the dataset has been preliminarily divided in in train set (70%)
 411 and a validation set (30%) and the evaluation is done separately on both. In fig.
 412 4 and fig. 6 the time dependant AUCs of the two survival models with respect
 413 to the train set are plotted. In fig. 5 and fig. 7 the same is done with respect to
 414 the validation set. Logistic regression accuracy is also plotted for reference with
 415 a non-survival type model, as mentioned above. The plots show the survival
 416 forest to be the better model. This is true both in terms of accuracy and of
 417 stability. In fact the AUCs corresponding to the random forest are higher than
 418 those of the AFT model. Further RF appears to be less unstable as the char-
 419 acteristics, such as dimension and proportion of surrenders, of the sub-datasets
 420 change. This is appears clearer when confronting figg. 4-7 with fig. 2 and fig.
 421 3, which represent the distribution of the trimesters. In particular, the greatest
 422 instability for RF happens at the least populated trimesters. This is true also
 423 for the AFT model with the difference that high instability is present also at
 424 the most populated trimesters. In other terms the survival regression model is
 425 stable only for moderately large sub-datasets. Note also that the logistic regres-
 426 sion behaves similarly to the random forest. It's characterized by relatively low
 427 stability for the initial times and convergence at later times, but tends to slowly
 428 decline. All of the considerations done so far are true both when considering
 429 the train set as when considering the validation set.

430 Having evaluated the models in terms of adherence with actual surrender fre-
 431 quency and of time-dependant AUC, we now add a final aggregated performance
 432 measure. This is very intuitively given by a pondered average of the AUC time
 433 series. The weights are given by the number of individuals present at the given
 434 time t , i.e. by the number of observations in the t^{th} sub-dataset. In formula:

$$AUC^D = \frac{\sum_{t=1}^N AUC_t^D n_t}{\sum_{t=1}^N n_t}, \quad AUC^P = \frac{\sum_{t=1}^N AUC_t^P n_t}{\sum_{t=1}^N n_t} \quad (21)$$

435 As summarized in tables 4-5 and illustrated in figg. 4-7, the survival forest
 436 performs better also in terms of the aggregated measure. The logistic regression
 437 appears to be more stable than the survival models as mentioned above but in
 438 terms of overall performance it places itself not only under the survival forest

	AFT	Forest
default	70.95%	93.56%
payoff	72.50%	90.60%

Table 4: Global AUC (21) for the surrender estimations (train set).

	AFT	Forest
default	72.07%	91.04%
payoff	73.40%	88.96%

Table 5: Global AUC (21) for the surrender estimations (validation set).

439 but also under the AFT model, with the only exception of default estimation on
440 train set, where the survival regression performs worse than the logistic. The
441 global AUC for default is equal to 71.43% while that for prepayment is equal to
442 64.37%.

443 The best performing model is thus the random survival forest, which can be
444 used to estimate the lapse rate. Interesting observations can be made on how
445 this quantity varies as it's influencing features vary. Among others, we're in-
446 terested in the effect of the interest rate, seeing as many empirical studies have
447 demonstrated the relevance of this variable (see Outreville (1990)). Note that
448 since the two options underlying the lapse rate estimation represent opposite
449 situations in terms of the policyholders economic availability and tend to be
450 influenced in opposite ways by the influencing features, we prefer to analyze
451 them separately. The fig. 8-31 thus represent default and prepayment condi-
452 tional probabilities and not lapse rate. In particular, we consider a baseline
453 policyholder profile by setting all influencing features at their average level with
454 regards to the reference dataset. In correspondence of this profile, we then es-
455 timate the baseline default and prepayment rates, which clearly vary with time
456 as do the survival curves from which they are computed. The baseline default
457 and prepayment rates are then confronted with those obtained with keeping all
458 other characteristic still and varying just one of them at a time. In particular,
459 we consider two extreme cases represented by the first and ninth decile of the
460 given feature.

461 With regards to the origination value of the interest rate applied to the loan, we
462 find that the default rate corresponding to the higher decile is mostly greater
463 than that corresponding to the average or lower decile, which instead almost
464 coincide. Note that these values correspond respectively to 4%, 6% and 9%.
465 Note also that the second is the contractual spread that we will use in the lapse
466 prediction in Section 6.3. Always with regards to interest rate, we repeat the
467 same analysis for `interest_rate_means` and find similar results. Non surpris-
468 ingly, having to pay a higher interest rate on the loan leads to higher default
469 rate.

470 Another interesting observation can be made by considering the logarithm of
471 loan value. In particular, we find that the higher the initial value of the loan
472 the higher the default rate. Note also that the categorical features (whether
473 the loan is take out for a single family house or a condominium, whether we're
474 talking about an area with planned urban development or whether the borrower

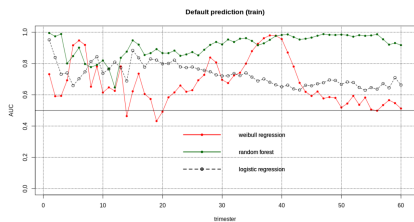


Fig. 4: $AUC(t)$ plot for default probability estimation (training)

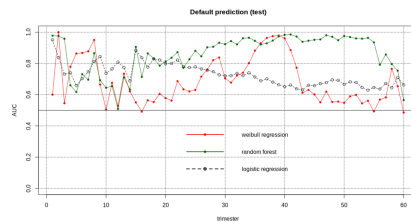


Fig. 5: $AUC(t)$ plot for default probability estimation (validation)

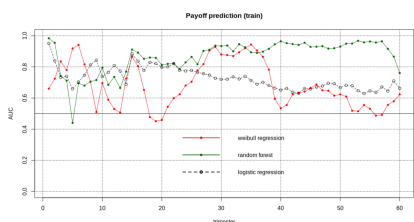


Fig. 6: $AUC(t)$ plot for prepayment probability estimation (training)

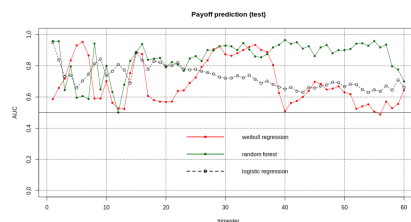


Fig. 7: $AUC(t)$ plot for prepayment probability estimation (validation)

475 is an investor or not), don't influence excessively the default rate. Further, with
 476 regards to the FICO credit score we find that the higher the credit score the
 477 lower the chance of default. While this is true globally it doesn't seem as un-
 478 ambiguous in the central times as it does elsewhere, this is probably due to
 479 the fact that the difference in the three credit scores considered are not as far
 480 apart. Finally note that with respect to the macroeconomic variables we find a
 481 rather extreme behavior when considering `uer_beta` and `gdp_beta` compared to
 482 what we've seen so far. In particular, a higher variation in unemployment rate
 483 over a unitary period of time, i.e. a trimester, leads globally to higher default
 484 rate. The opposite is true for a higher variation in gross domestic product over
 485 a trimester, which in fact leads to significantly lower default rate. Both of these
 486 observations are quite intuitive.

487 When considering prepayment, instead, we find that higher chance of prepay-
 488 ment corresponds to lower average interest rate. Further, non surprisingly ex-
 489 cellent FICO credit score lead to high chance of prepayment. This is true in
 490 more unambiguous way compared with what we saw for the default probability.
 491 Another clear and interesting result is that the gross domestic product value
 492 at the origination time of the mortgage as a noticeable impact on prepayment
 493 probability. In particular, the high the value of `gdp_beta` the higher the chance
 494 of prepayment. Note instead that the prepayment probability we find with the
 495 first decile of the given feature is quite similar to what we find for the average
 496 value. Interestingly, the effect of the origination value of the gross domestic
 497 product on the chance is prepayment is much less unambiguous then when con-
 498 sidering chance of default (see fig. 12 and fig. 24). This seems to confirm how
 499 the prepayment option is more linked to personal conditions and preference then
 500 is default, which instead depends more on the characteristics of the contract.

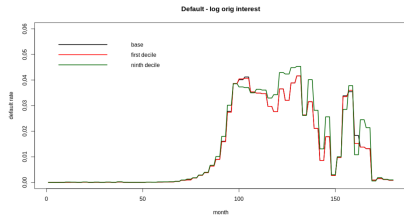


Fig. 8: Default rate at base value, 1st and 9th decile of interest_rate_orig.

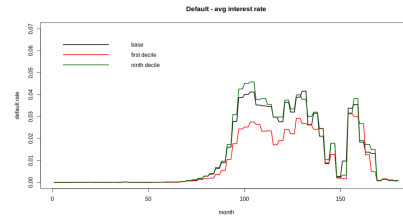


Fig. 9: Default rate at base value, 1st and 9th decile of interest_rate_means.

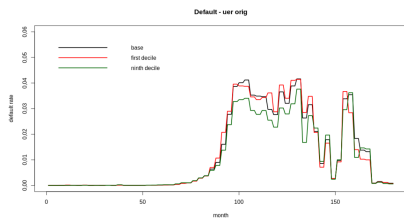


Fig. 10: Default rate at base value, 1st and 9th decile of uer_orig.

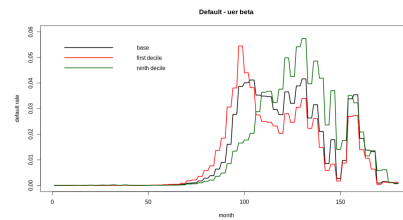


Fig. 11: Default rate at base value, 1st and 9th decile of uer_beta.

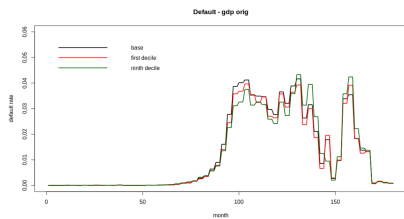


Fig. 12: Default rate at base value, 1st and 9th decile of gdp_orig.

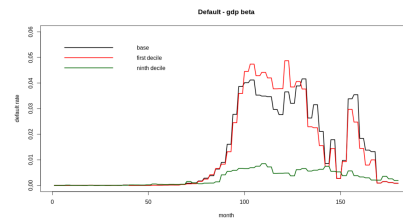


Fig. 13: Default rate at base value, 1st and 9th decile of gdp_beta.

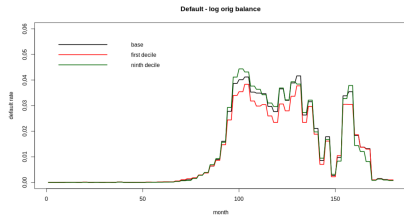


Fig. 14: Default rate at base value, 1st and 9th decile of log_balance_orig.

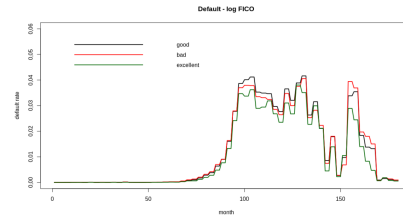


Fig. 15: Default rate at good, bad and excellent level of log_FICO_orig.

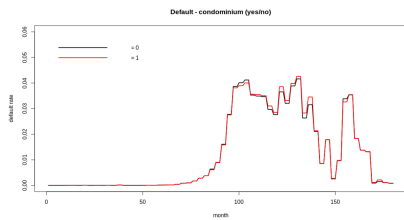


Fig. 16: Default rate for yes/no value of RType_CO_orig.

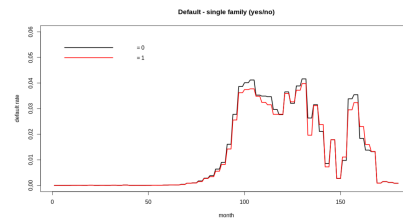


Fig. 17: Default rate for yes/no value of RType_SF_orig.

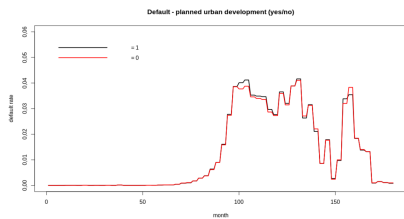


Fig. 18: Default rate for yes/no value of RType_PU_orig.

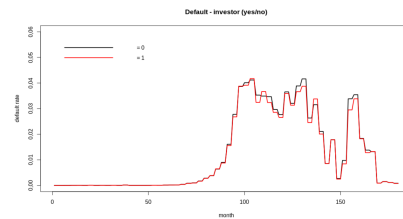


Fig. 19: Default rate for yes/no value of investor_orig.

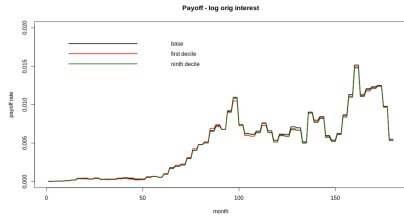


Fig. 20: Prepayment rate at base value, 1st and 9th decile of `interest_rate_orig`.

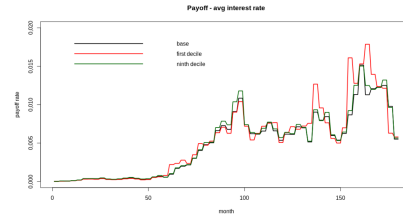


Fig. 21: Prepayment rate at base value, 1st and 9th decile of `interest_rate_means`.

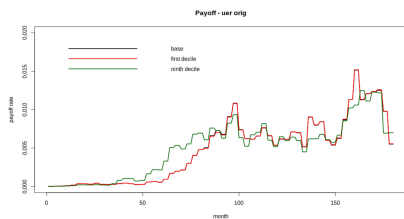


Fig. 22: Prepayment rate at base value, 1st and 9th decile of `uer_orig`.

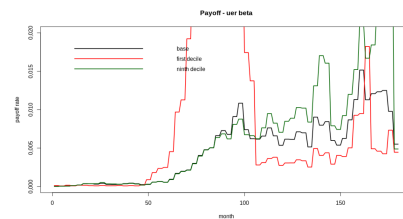


Fig. 23: Prepayment rate at base value, 1st and 9th decile of `uer_beta`.

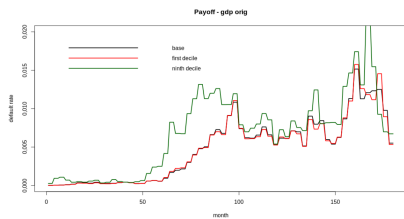


Fig. 24: Prepayment rate at base value, 1st and 9th decile of `gdp_orig`.

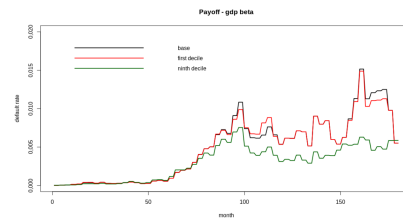


Fig. 25: Prepayment rate at base value, 1st and 9th decile of `gdp_beta`.

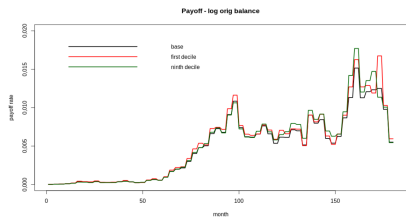


Fig. 26: Prepayment rate at base value, 1st and 9th decile of log_balance_orig.

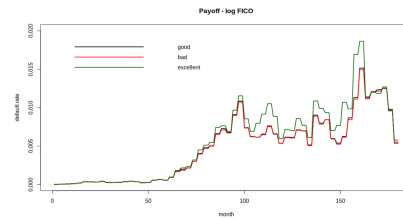


Fig. 27: Prepayment rate at good, bad and excellent level of log_FICO_orig.

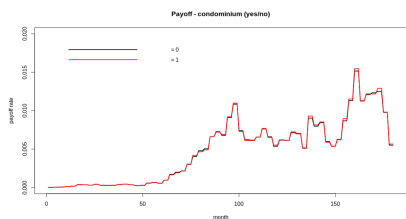


Fig. 28: Prepayment rate for yes/no value of REtype_CO_orig.

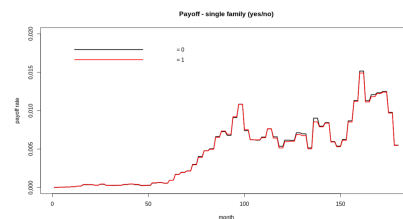


Fig. 29: Prepayment rate for yes/no value of REtype_SF_orig.

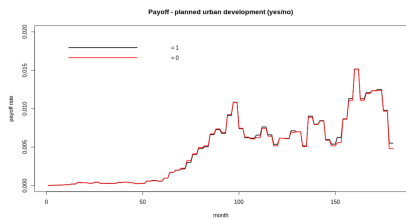


Fig. 30: Prepayment rate for yes/no value of REtype_PU_orig.

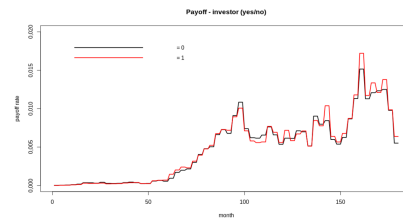


Fig. 31: Prepayment rate for yes/no value of investor_orig.

a	b	σ
0.2	0.005	0.0005

Table 6: Reference Vasicek parameters.

501 A final general observation can be done for default and prepayment probabilities:
 502 they begin to become relevant from the thirtieth trimester onward, while
 503 before they are very close to zero, independently of whether we are considering
 504 default or prepayment, intermediate or extreme feature values. This is proba-
 505 bly due to the fact that before the thirtieth trimester the number of observed
 506 policyholder is quite low.

507 6.3 TVOG analysis

508 Now that we have obtained the surrender probabilities and that we have ade-
 509 quately evaluated and confronted our models, we can integrate the estimated
 510 lapse rate in the CPI model. Note that the estimate is that of the best per-
 511 forming model, i.e. random survival forest, which hereafter will be the reference
 512 model. The profit and the premium are determined in three steps:

- 513 1. simulation of the interest rate scenarios through the model mentioned in
 514 Section 3
- 515 2. lapse rate estimation through the fitted model
- 516 3. profit and premium determination through a cash flow model that takes
 517 lapse rate, exogenously set quantities and interest rate scenarios in input

518 With regards to the first step, we must preliminarily make a distinction between
 519 the base scenario, i.e. the certain equivalent scenario, and the complete stochas-
 520 tic scenarios. The first doesn't comprise the Brownian motion component and
 521 is thus not subject to the volatility in the results we find in the second. We in
 522 fact simulate $N = 1000$ complete stochastic interest rate scenarios with enough
 523 volatility to appreciate a distinction from the certain equivalent scenario. As
 524 a consequence the result may vary sensibly between each set of N simulations.
 525 This variability can be absorbed by increasing the number of simulations or
 526 setting a seed, thus rendering the results comparable.

527 As mentioned in section 3, the parameters of the Vasicek model aren't calibrated
 528 on the historical market rights, but are set exogenously. In table 6 the reference
 529 parameters. With these settings we simulate $N = 1000$ forward curves and one
 530 certain equivalent interest rate curve. This is done for the loan duration of 180
 531 months, i.e. 15 years. With regards to the second step, we consider a policy-
 532 holder with an average profile compared to the reference dataset. The initial
 533 value of the loan is 100.000 and the rate at which the capital is unmounted, is set
 534 to 6%. The policyholder's credit score is good, as per the official FICO credit
 535 scoring system ($\log_FICO_orig = \log(700)$), and we consider unemployment
 536 rate and gross domestic product equal to the average over the period covered
 537 by the dataset. In particular, the origination unemployment rate is set at 5.1%
 538 while the origination gross domestic product rate is 2.9%. Further we have an
 539 upward unemployment trend over the considered period of 0.0005 per unit of

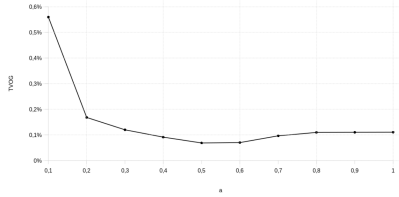


Fig. 32: TVOG shape by a .

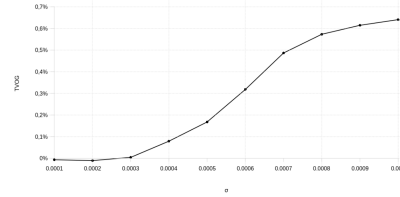


Fig. 33: TVOG shape by σ .

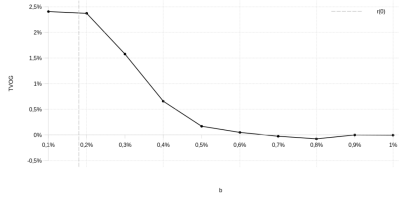


Fig. 34: TVOG shape by b .

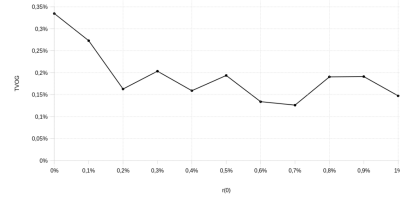


Fig. 35: TVOG shape by $r(0)$.

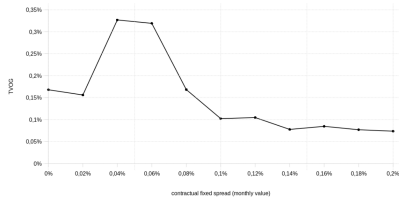


Fig. 36: TVOG shape by s .

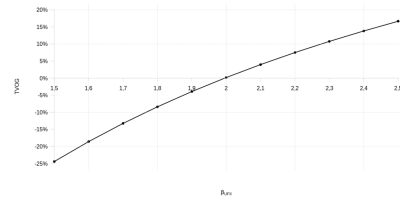


Fig. 37: TVOG shape by β .

540 time. With regards to gross domestic product we have a downward trend of -
 541 0.001 per unit of time. Note also that the simulated forward curves are monthly
 542 projections while the interest rates on which the model is fitted are quarterly.
 543 Once this has been adequately accounted for, we obtain the estimated survival
 544 curve, the conditional surrender probabilities (14) and the lapse rate (15).
 545 The third and final step consists in integrating the simulated forward curves
 546 and the estimated lapse rate into a cash-flow model. As explained in Section 3,
 547 the amortization rate paid is a linear function of the simulated interest rates.
 548 In particular, the contractual spread is set to 0%, and the simulated forward
 549 curves are multiplied by coefficient set to 1. Further, it's minimized at 1%, i.e.
 550 0.083% monthly. Also the investment rate is defined as a linear function of the
 551 simulated forward curve. The spread is set at 0% and the interest rates are mul-
 552 tiplied by a parameter set equal to 2. Taxes and expenses are annulled. With
 553 these settings we obtain certain equivalent profit of 2071.25 and a complete
 554 stochastic profit of 2067.77. The proportion with the premium is respectively
 555 68.6% and 68.5%. Finally TVOG is equal to 0.17%.
 556 Interesting conclusions can be drawn by varying one of the parameters that
 557 enter the determination of TVOG. In particular we will consider TVOG as a
 558 function of the three Vasicek parameters, of the initial risk-free rate, of the

559 contractual spread applied to in order to determine installments and of the invest-
560 ment parameter. In short $TVOG = f(a, b, \sigma, r(0), spread_{INSTALL}, \beta_{UFII})$.
561 With regards to a we find that the baseline value corresponds to and elbow in
562 the TVOG distribution. A first steep decrease is found when passing from 0.1
563 convergence velocity to 0.2 while the following values imply a level of TVOG
564 which stabilized around 0.1%. This is not surprising as we expect its value
565 to converge to a certain level as a becomes large enough to make the forward
566 curve align to the average level $b = 0.005$ sooner and sooner. The volatility is
567 instead varied between 0.0001 and 0.001 and the corresponding TVOG curve
568 has an increasing monotone form with an inflection point around the baseline
569 σ value. This means that beyond this level of volatility TVOG increases at de-
570 creasing pace but nonetheless continues increasing. Further with regards to the
571 average level b we find that TVOG increases for values smaller than the initial
572 risk-free rate $r(0)$ and begins decreasing for larger values, eventually converging
573 0%. The shape of the TVOG curve for values smaller than $r(0)$ shouldn't come
574 as a surprise as it depends on the structure of the Vasicek differential equation.
575 Clearly the same is true when considering different values for the initial risk-free
576 rate. Note that the baseline initial risk-free interest rate is set equal to the last
577 of the monthly overnight LIBOR time series considered. This value is near to
578 the elbow of the distribution which decreases initially to then stabilize within
579 a band of values 0.1% - 0.2%. With regards to the contractual spread applied
580 to the installments we find that the TVOG converges to just under 1% after an
581 initial steep rise. Finally the dependence of TVOG from the β applied to the
582 investments is quite smooth. Note that the spread applied to the investment
583 rate is equal to zero. This means that the rate itself is merely a multiple of
584 the overnight LIBOR. When the coefficient is low, i.e. $\beta_{UFII} < 2$, TVOG is
585 negative. The higher the coefficient the higher the TVOG. Note also that this
586 happens quite rapidly as in a span of -1.5 to 2.5 β_{UFII} TVOG passes from
587 -25% to 20%.

588 6.4 Limitations, extensions and conclusions

589 TODO

590 References

- 591 B. Baesens, D. Roesch, H. Scheule, *Credit Risk Analytics: Measurement Tech-*
592 *niques, Applications and Examples in SAS*, Wiley, 2016.
- 593 C. F. B. Richardson, J. M. Hartwell, *Lapse rates*, Transactions of Society of
594 Actuaries, 1951, Vol. 3 No. 7.
- 595 J. F. Outreville, *Whole-life insurance lapse rates and the emergency fund hy-*
596 *pothesis*, Insurance: Mathematics and Economics, 9(4), 249-255.
- 597 Sv. Nolte, Sc. Nolte, C. Judith, *Dont lapse into temptation: a behavioral ex-*
598 *planation for policy surrender*, Journal of banking and finance, 2017, 79:12-27.
- 599 F. Barsotti, X. Milhaud, Y. Salhi, *Lapse risk in life insurance: Correlation and*
600 *contagion effects among policyholders behaviors*, Insurance: Mathematics and
601 Economics, 2016, Vol. 71, 317-331.
- 602 M. Li, *Residential mortgage probability default models and methods*, Risk
603 Surveillance and Analytics, Financial Institution Commission.
- 604 D. R. Capozza, D. Kazarian, T. A. Thomson, *The conditional probability of*
605 *mortgage default*, 1996, Real Estate Economics 26-3: 359-389.
- 606 *Consultation paper on the proposal for guidelines on Own Risk and Solvency*
607 *Assessment*, 7 November 2011, EIOPA-CP-11/008.
- 608 S. Loisel, X. Milhaud, *From deterministic to stochastic surrender risk models:*
609 *Impact of correlation crises on economic capital*, 2011, European Journal of
610 Operational Research, Vol. 214, 348-357.
- 611 J. C. Hull, *Options, futures and other derivatives*, Pearson, 8th edition, 2012.
- 612 H. Ishwaran, U. B. Kogalur, *Random Survival Forest for R*, R News Vol. 7/2,
613 2007.
- 614 H. Ishwaran, U. B. Kogalur, E. H. Blackstone, M. S. Lauer, *Random survival*
615 *forests*, Annals of Applied Statistics 3, 841-860, 2008.
- 616 M. LeBlanc, J. Crowley, *Relative risk trees for censored survival data*, Biomet-
617 rics 48, 411-425, 1992.
- 618 T. Therneau, P. Grambsch, T. Fleming, *rpart: Recursive Partitioning*, R pack-
619 age version 3.1-46, 2010.
- 620 A. N. Kamarudin, T. Cox, R. Kolamunnage-Dona, *Time-dependent ROC curve*
621 *analysis in medical research: current methods and applications*, BMC Medical
622 Research Methodology, BMC series, 17:53, 2017.