MODELING LAPSE RATES USING MACHINE LEARNING: A COMPARISON BETWEEN SURVIVAL FORESTS AND COX PROPORTIONAL HAZARDS TECHNIQUES

MODELIZACIÓN DE TASAS DE CAÍDAS UTILIZANDO MACHINE LEARNING: COMPARACIÓN ENTRE LAS TÉCNICAS DE SURVIVAL FOREST Y COX PROPOTIONAL HAZARDS

Jorge Luis Andrade[1], José Luis Valencia

Facultad de Estudios Estadísticos. Universidad Complutense de Madrid
Avenida Puerta de Hierro s/n Ciudad Universitaria 28040. Madrid

Fecha de recepción: 13/04/2021
Fecha de aceptación: 14/07/2021

Abstract

This study undertakes a comparative analysis of the performance of machine learning and traditional survival analysis techniques in the insurance industry. The techniques compared are the traditional Cox Proportional Hazards (CPH), Random Survival Forests (RSF) and Conditional Inference Forests (CIF) machine learning models. These techniques are applied in a case study of insurance portfolio of one of Ecuador's largest insurer. This study demonstrates how machine learning techniques perform better in predicting survival function measured by the C-index and Brier Score. It also demonstrates that the predictive contribution of covariates in the RSF model is consistent with the traditional CPH model.

Keywords: análisis de supervivencia, machine learning, tasas de caídas, random survival forest.

Resumen

Este estudio realiza un análisis comparativo del rendimiento de las técnicas de machine learning y tradicionales de análisis de supervivencia. Las técnicas

[1] Autor para correspondencia: jorandra@ucm.es

DOI 10.26360/2021_7

comparadas son el tradicional modelo de Cox Proportional Hazards (CPH) y las técnicas de machine learning Random Survival Forest (RSF) y Conditional Inference Forest (CIF). Estas técnicas se aplican para el estudio de una cartera de seguros de una de las Compañías más grandes de Seguros de Ecuador. Este estudio demuestra un mejor rendimiento de las técnicas de machine learning en la predicción de la función de supervivencia medidos a través del C-index y el Brier Score. También se demuestra que la aportación predictiva de las covariables en el modelo RSF es consistente con el modelo tradicional CPH.

Palabras clave: análisis de supervivencia, machine learning, tasas de caídas, random survival forest.

1. Introduction

Under the global regulatory framework of IFRS 17 (International Accounting Standards Board, 2017) and the European Solvency II standard, insurance companies are required to estimate future lapse rates to calculate their best estimate liabilities.

According to EIOPA (2010) the lapse risk in an insurance portfolio is the most important risk, accounting for almost 40% of the capital requirement within the *life underwriting risk* module that includes risk factors such as longevity, mortality, disability, catastrophe and expenses.

Modeling the risk of lapse rates in life insurance is directly related to the benefit derived from the insured lapse. This benefit results from the penalty on the provision made at the date of lapse. A penalty is set over the lapsed amount to compensate for the increase in mortality risk in the remaining portfolio, to avoid a financial mismatch in insurance cash flows, and to discourage a lapse decision by policyholder (Eling and Kiesenbauer, 2014; Eling and Kochanski, 2013). Under such conditions, high lapse rates decrease insurance reserves but increase capital requirements.

Traditional Cox proportional hazards (CPH) models have been generally used for estimating the survival function of the time-to-event random variable and, consequently, the future lapse rates (Brockett et al., 2008; Eling and Kiesenbauer, 2014; Pinquet, Guillén, and Ayuso, 2011), though more recently, a few studies have used machine learning techniques (Aleandri and Eletti, 2020).

Predictive modeling was proposed in Eling and Kochanski (2013) to explore the problem. Machine learning predictive techniques, such as random survival forest (RSF) present advantages in that they estimate future lapse rate probabilities at the policy level without assuming a proportional relationship among

policyholders, nor assuming an identical base function for all policyholders. Furthermore, it's robust to outliers and does not suffer from convergence problem. In this study, the RSF, conditional inference forests (CIF), and CPH models are applied at the policy level in the insurance portfolio of an Ecuadorian life insurance company. A comparative analysis is undertaken to evaluate the viability of machine learning techniques, by interpreting their outputs and assessing their consistency. The study concludes by providing some practical actuarial considerations.

For the development of these models, the following R libraries were selected: *randomForestSRC* (Ishwaran and Kogalur, 2017), *survival* Therneau and Lumley, 2015), *party* (Zeileis, Hothorn, and Hornik, 2010) and *PEC* (prediction error curves) (Gerds, 2017); and the outputs were compared to their equivalents in the *ranger* package using C++ code (Wright and Ziegler, 2017) and also implemented in *PySurvival*} Python package (Fotso et al., 2019).

The models section outlines the theoretical formulation and algorithms for estimating the survival function, then provides two indicators to assess the predictive power of the models.

The data description section of this paper describes and analyzes the dataset.

The analysis and results section compares the outcomes of the CPH and RSF models, and the importance of covariates in the lapses is explained. For practical actuarial purposes, the RSF model allows for the prediction of individual lapse rates and the creation of differentiated risk groups. The final section compares the predictive power of the proposed models.

2. Models

This section details the mathematical formulation and the algorithms of the various models used for survival analysis –namely, the machine learning techniques RSF and CIF, and the traditional CPH model.

The survival function of the random variable T is given by $S(t) = Pr(T \geq t)$. For the modeling of T, it is necessary to set $(T_i, \delta_i, \mathbf{X}_i)$, where $i \in 1,..,n$ are individuals, T_i is the time-to-event, δ_i is as defined by (1), and $\mathbf{X}_i = (x_{1i},...,x_{pi})$ is the vector of explanatory variables.

$$\delta_i = \begin{cases} 1 \text{ if status=lapse} \\ \\ 0 \text{ otherwise.} \end{cases} \quad (1)$$

2.1 Random Survival Forests (RSF)

Random Forests is one of the most interesting machine learning techniques for classification and regression. This technique was applied to survival analysis in a proposal of Ishwaran, Kogalur, Blackstone, and Lauer (2008) as an extension of the original paper of Breiman (2001).

An advantage of this model is that it is totally non-parametric, and therefore, does not assume a distribution for the relationship between the predictors and the response variable. In addition, it captures linear and non-linear relationships between the explained variable and the predictor variables. Another important feature is that it finds interactions between covariates because the learning comes from the ensemble decision trees. Outliers in data does not affect it, nor suffer from convergence problems.

This model does not require the assumption of the CPH model that there are proportional hazards among individuals; instead, it allows the construction of survival functions with different shapes for each insured. In addition, the assumption of the same hazard rate basic for all insured is avoided because it is inconsistent with the reality.

Survival trees are built by splitting each parent node into two daughter nodes starting at the root, which comprising the full dataset. A split is performed according to a survival criterion that maximizes the difference between daughter nodes (we use the log-rank test explained below); such a split is repeated on each subsequent node in a binary manner. This process is repeated to build *n* trees, and then ensemble techniques are used to obtain the final estimators, in this case the average of all trees.

The algorithm has double randomness. First, a random sample is obtained by replacing the original data in each new tree. Second, the parent node is split into two daughters using a randomly selected covariate x_j. Due to the *law of large numbers*, this double randomness leads to the convergence of the prediction error (PE) (refer to Figure Figure 1). It describes the number of trees larger the PE converge with a higher accuracy Breiman (2001).

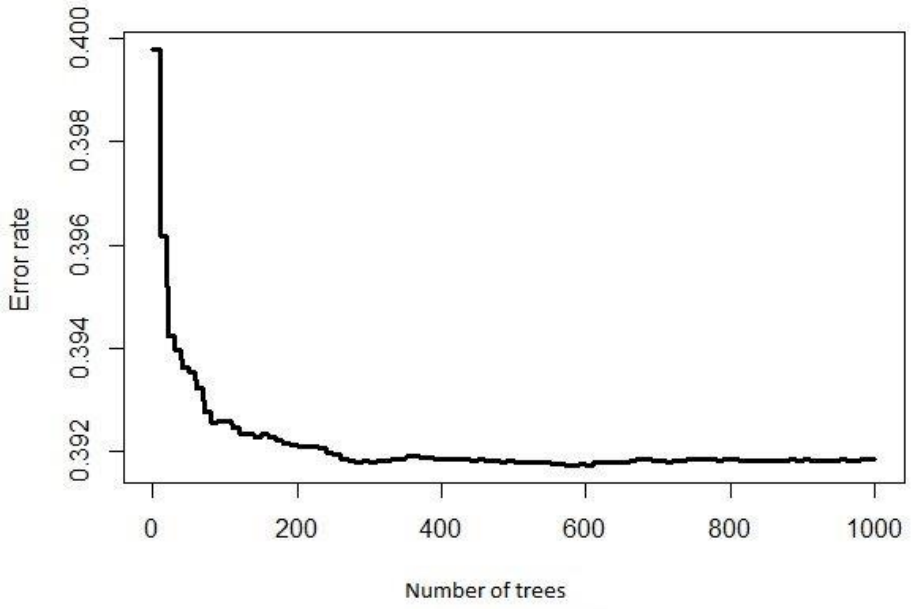

Figure 1. Demonstration of prediction error convergence using our dataset

The algorithm is implemented as follows:

1. Draw B samples of average size comprising 63% of the original data with replacement (average bootstrap sample in *n* trials -1/*e* approx).
2. Grow a survival tree for each bootstrap sample; at each tree node, *p* covariates are randomly selected.
3. The parent node is split with respect to covariate x^* and its value c^*, which maximizes the survival difference in the score function as defined in equation (3).
4. Expand the tree to its maximum size provided that the terminal nodes have $d > 0$ event cases.
5. Calculate the cumulative hazard function (CHF) for each terminal node of the tree, as defined in equation (5).
6. Average the CHF and survival function over the bootstrap samples at terminal nodes.
7. Calculate the PE of the ensemble according to C-index described in section below.

In this study, the survival criterion used to make the split of each parent node *w* is the log-rank. The cutoff c_j for each variable x_j is determined in such a way that one daughter node maintains the values $x_{ji} \le c_j$ and the other maintains the observations $x_{ji} > c_j$:

$$Y_{k,l}^{j} = \{i : T_i \ge t_k, x_{ji} \le c_j\}, Y_{k,r}^{j} = \{i : T_i \ge t_k, x_{ji} > c_j\}, \tag{2}$$

where x_{ji} is the value of x_j for individual $i = 1,...,n$, $d_{k,l}^{j}$ and $Y_{k,l}^{j}$ are the events and the policies exposed at risk respectively in the daughter left node l.
Similarly let $d_{k,r}^{j}$ and $Y_{k,r}^{j}$ refer to the right daughter node, let $Y_k = Y_{k,l}^{j} + Y_{k,r}^{j}$ and $d_k = d_{k,l}^{j} + d_{k,r}^{j}$; and *n* is the number of insured at the parent node under condition that $n = n_{k,l}^{j} + n_{k,r}^{j}$.

Let $t_1 < t_2 < ... < t_m$ be the different event times in the parent node *w*, the log-rank test for a split at the value c_j for an x-variable x_j is:

$$L(c_j, x_j) = \left(\frac{\sum_{k=1}^{m} (d_{k,l}^{j} - Y_{k,l}^{j} \frac{d_k}{Y_k})}{\sqrt{\sum_{k=1}^{m} \frac{Y_{k,l}^{j}}{Y_k}(1 - \frac{Y_{k,l}^{j}}{Y_k})(\frac{Y_k - d_k}{Y_k - 1}) d_k}} \right) \tag{3}$$

The choice of the best split for *w* is determined by x^* and maximizing the difference in survival criterion between the daughter nodes $\{r, l\}$ such that $|L(c^*, x^*)| \ge |L(c_j, x_j)|$ for all *j*.

A terminal node *h* is found when a saturation point is reached because no new daughters can be formed.

Once *B* trees are obtained, the ensemble technique is applied. At each node terminal *h* in all trees, the cumulative hazard function (CHF) is calculated with the Nelson-Aalen estimator over the $t_k \in \{t_1, t_2, t_3, ..., t_N\}$ and the insured status. All insured forming the terminal node *h* have the same CHF, then $H_b(t \mid \mathbf{X}_i) = H_h(t)$ if $\mathbf{X}_i \in h$,

$$H_h(t) = \sum_{t_k \le t} \frac{d_{k,h}}{Y_{k,h}} \tag{4}$$

An average over the bootstrap samples is calculated to obtain the CHF estimation $H_e(t \mid \mathbf{X}_i)$ in the forest. For an insured with covariates $\mathbf{X}_i$, $H_b(t \mid \mathbf{X}_i)$ is the CHF, where *B* is the number of samples:

$$H_e(t \mid \mathbf{X}_i) = \frac{1}{B} \sum_{b=1}^{B} H_b(t \mid \mathbf{X}_i) \tag{5}$$

The non-parametric Kaplan-Meier is used to predict the survival function in the forest. For an insured with covariates $\mathbf{X}_i$ and *B* samples, it is:

$$S_e(t \mid \mathbf{X}_i) = \frac{1}{B} \sum_{b=1}^{B} \left(\prod_{t_{k,b} \le t} \frac{(Y_{t_{k,b}} - d_{t_{k,b}})}{Y_{t_{k,b}}} \right) \tag{6}$$

The RSF model exhibits the conservation-of-events principle, which asserts that the sum of the estimated CHF over the observed time at each terminal node in a tree equals the total number of events in the dataset.

2.2 Conditional Inference Forests (CIF)

A CIF algorithm is proposed by Hothorn, Hornik, and Zeileis (2006). This presents advantages similar to the RSF model in comparison with traditional models, but the split calculation and the ensemble formula are different.

The algorithm is implemented as follows:

1. Test the null hypothesis of independence between any of the *p* covariates and the response variable; the test is based on log-rank transformed statistic.
2. Tree growth is stopped if the null hypothesis cannot be rejected; otherwise, a x_j^{th} covariate with the strongest level is selected according its *p*-value.
3. The covariate selected x_j is divided into two disjoints sets based on multiples adjusted *p*-values (Montecarlo simulations or Bonferroni corrections), on univariates *p*-values or on values of the test statistic; and when the specified criterion is exceeded, the parent node is split.

4. Apply the ensemble to the samples to obtain the survival function(explained below) and the CHF.

The survival function CIF ensemble is a Kaplan-Meier weighted estimator:

$$S_e(t \mid \mathbf{X}_i) = \prod_{t_{k,b} < t} \left(1 - \frac{\sum_{b=1}^{B} d_{k,b}}{\sum_{b=1}^{B} Y_{k,b}} \right) \tag{7}$$

2.3 Cox Proportional Hazards (CPH)

In this section, the survival function $S(t \mid \mathbf{X}_i)$ is adjusted by applying the classic Cox model technique using the assumption of proportional hazards because this is one of the most widely used in survival analysis. The adjustment is made taking into account the model assumptions and using the linear estimation model for the *log* of the hazard rate $logh_i(t \mid \mathbf{X}_i)$ as per Crumer (2011), based on an exponential distribution:

$$logh_i(t \mid \mathbf{X}_i) = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi} \tag{8}$$

the insured hazard rate with a vector of covariates $\mathbf{X}_i$ is defined:

$$h_i(t \mid \mathbf{X}_i) = exp(\alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi}) \tag{9}$$

where i is an insured, α is a constant which represents the log-base hazard $logh_o(t)$, and $\beta_1 \ldots \beta_p$ are the parameters estimated (β) using the partial likelihood.

The partial likelihood function is the product of the conditional probability for all individuals in the sample according to Cox's proporsal. It does not involve the unspecified underlying hazard function h_o but covariates, and β coefficients that can be estimated. Censored observations contribute to the partial likelihood through the risk sets.

Suppose that m of the survival times from n individuals are uncensored and distinct, and $n - m$ $t_1 < t_2 < \ldots < t_m$ be the ordered m

$R_{t(i)}$ be the risk set at time t_i, it consists of all insured who are at risk at time t_i. The probability that the failure is on the individual *i* at time $t_{(i)}$ is $P(i,t_{(i)})$:

$$P(i,t_{(i)}) = \frac{exp(\sum_{j=1}^{p} \beta_j x_{j(i)})}{\sum_{l \in (R_{t(i)})} exp(\sum_{j=1}^{p} \beta_j x_{jl})}, \tag{10}$$

each lapse contributes a factor, and hence the partial likelihood function is:

$$L(\beta) = \prod_{i=1}^{m} \frac{exp(\sum_{j=1}^{p} \beta_j x_{j(i)})}{\sum_{l \in (R_{t(i)})} exp(\sum_{j=1}^{p} \beta_j x_{jl})} \tag{11}$$

After applying the log-likelihood function, β coefficients are estimated solving the simultaneous equations using iterative procedures like Newton-Raphson (for more details see Lee and Wang (2013)).

When all covariates are set to zero $\alpha(t) = logh_o(t)$, and the log-hazard function is defined by:

$$h_i(t \mid \mathbf{X}_i) = h_o(t) exp(\beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_p x_{pi}), \tag{12}$$

this leads to the survival function equation:

$$S_e(t \mid \mathbf{X}_i) = exp(-h_o(t) exp(\beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_p x_{pi})). \tag{13}$$

To demonstrate the proportional hazard assumption in the Cox model, two insured *a* and *b* are considered. Let $\phi_a = \beta_1 x_{1a} + \beta_2 x_{2a} + ... + \beta_p x_{pa}$ and $\phi_b = \beta_1 x_{1b} + \beta_2 x_{2b} + ... + \beta_p x_{pb}$. The ratio of their estimated hazard is provided by:

$$\frac{h_a(t)}{h_b(t)} = \frac{h_o(t) exp(\phi_a)}{h_o(t) exp(\phi_b)} = exp[\phi_a - \phi_b]. \tag{14}$$

2.4 Assessment of predictive power

The performance of the proposed models were compared using two indicators ─the Concordance index (C-index) and the Brier Score─ which were measured over the test dataset. A model has a better predictive power when its C-index is higher, and its Brier Score is lower, than that of the other models.

To assess the **C-index**, an evaluation pair with different observed responses is said to be concordant if the observation with the lowest level of the response has a lower predicted value than the observation with the highest level.

The C-index evaluates for each insured the performance of the model at each time *t*, and it measures how well the predictor ranks two randomly selected insured lives in terms of survival.

In survival analysis, the predictive power of the model is measured by C-index, and is equivalent to the area under of curve (AUC). In addition, the error rate for a survival model is measured by 1-C-index.

Let $t \in \{t_1, t_2, t_3, ..., t_m\}$ denote each unique times in the test dataset. The predicted C-index for insured is defined by:

$$\Gamma_{(i)} = \sum_{k=1}^{m} H_e(t_k \mid \mathbf{X}_i), \tag{15}$$

an individual *i* has a worse predicted outcome than individual *j* if $\Gamma_{(i)} > \Gamma_{(j)}$.

The C-index is calculated as follows (Ishwaran and Kogalur, 2017):

1. Form all possible pairs of observations over all data.
2. Omit those pairs where the shorter event time is censored. In addition, omit pairs (i, j) if $T_i = T_j$ unless $(\delta_i = 1, \delta_j = 0)$ or $(\delta_i = 0, \delta_j = 1)$ or $(\delta_i = 1, \delta_j = 1)$. The last restriction only allows ties in event times if at least one of the observations is an event. Let the resulting pairs be denoted by Y. Let *Permissible* = $|\mathrm{Y}|$.
3. If $T_i \neq T_j$, count 1 for each $y \in \mathrm{Y}$ in which the shorter time had a worse predicted outcome.
4. If $T_i \neq T_j$, count 0.5 for each $y \in \mathrm{Y}$ in which $\Gamma_{(j)} = \Gamma_{(i)}$.

5. If $T_i = T_j$, count 1 for each $y \in \mathrm{Y}$ in which $\Gamma_{(j)} = \Gamma_{(i)}$.
6. If $T_i = T_j$, count 0.5 for each $y \in \mathrm{Y}$ in which $\Gamma_{(j)} \neq \Gamma_{(i)}$.
7. The C-index is calculated count over all permissible pairs. Final ratio is calculated by *C-index = Concordance/Permissible*.
8. The error rate is *Error*=1-C-*index*. If *C-index* = 0.5= *Error* this means that the model is doing no better than random guessing.

The second indicator to assess the performance of three models is the **Brier Score** $BS(t, S_e)$. It is defined as the error between the prediction and the actual data. For non-censored data, the formula is defined in equation (16) according to Mogensen, Ishwaran, and Gerds (2012):

$$BS(t, S_e) = E(Y_i(t) - S_e(t \mid \mathbf{X}_i))^2 \tag{16}$$

where $Y_i(t)$ is the observed status (0,1) of an individual *i* at time *t*.

For right censored data, the squared residuals for an insured are weighted using the inverse probability of censoring weights. The Brier Score integrated indicator for a dataset *D* of size *n* is calculated by:

$$BS(t, S_e) = \frac{1}{n} \sum_{i=1}^{n} \left\{ [0 - S_e(t \mid \mathbf{X}_i)]^2 \frac{I(t_i \leq t, \delta_i = 1)}{\hat{G}(t_i \mid \mathbf{X}_i)} + [1 - S_e(t \mid \mathbf{X}_i)]^2 \frac{I(t_i > t)}{\hat{G}(t \mid \mathbf{X}_i)} \right\} \tag{17}$$

where $\hat{G}(t \mid x) \approx P(C_i > t \mid \mathbf{X}_i = x)$ is the Kaplan-Meier estimate of the conditional survival function of the censoring times.

3. Data description

The data were derived from 39,572 policies issued by Ecuador's largest life insurer, covering a period of 15 years from 2005 to 2019.

The response and explanatory variables were processed and are described in Tables 1 and 2. All data processing was performed in Python.

In survival analysis, the response variables comprise the following variables: status, and time-to-event.

The company principally sells two individual life products—Universal Life and Term Life —both of which are well known in the actuarial literature. Term life insurance is a pure risk product that guarantees a capital benefit upon the death of the insured. Universal life insurance similarly guarantees a capital benefit upon the death of the insured, but also accumulates a cash value at a minimum interest rate and the policyholder may voluntarily opt for the lapse amount instead; the lapse amount is equal to the policy value less any penalty, which decreases with the age of the contract.

Table 1
List of categorical our dataset

Variable	Statistics	Comments
Status	In force 36,6%, Lapse 62,5%, others 0,95%	Others include: death and expiration
Sex	F 43,03%, M 56,97%	
Risk state	Smoker 6,37%, Non Smoker 93,63%	
Payment frequency	Annual 19,1%, Monthly 77,5%, Others 3,40%	Others include: four-monthly, quarterly and biannual
Product type	Term life 24,05%, Universal life 75,05%	The original data present different commercial names for both
Point of sales	Quito 36,9%, Guayaquil 47,1%, Cuenca 8,2%, Manta 7,1%, Ambato 0,7%	
Distribution channel	Corporate agents 11,7%, Individual agents 8,5%, Tied agents 79,8%, Others 0%	
Payment method	Bank debit 82,7%, Credit card 17,3%, Payroll 0%	
Profession	A 76,7%, B 18,7%, C 3,7%	Reserved variable

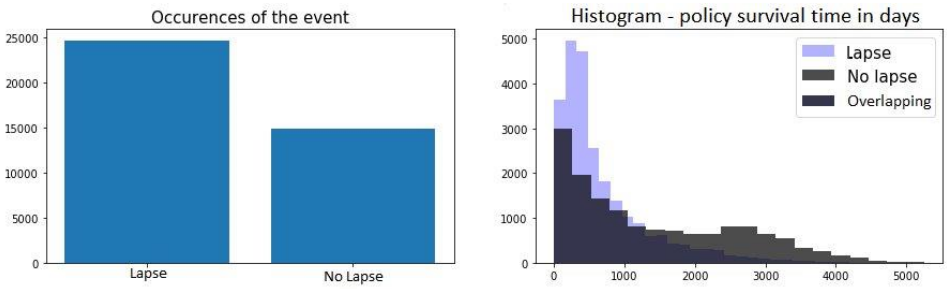

Figure 2. 1a) Status response variable, and 1b) Distribution of policies by policy survival time and status

The data are *right-censored* in type. There are three reasons for policies to be censored: policies that reach maturity, policies that at the date of the concluding the study had not experienced the event, or policies that during the study were terminated for reasons other than lapse.

Status is a binary variable that refers to the cancellation of an insurance contract before maturity. Lapse is the event of interest in this study. The other causes of termination are the policyholder's death, or policy expiration (premature termination guaranteed in law). The status variable was determined by equation (1) and its composition in our dataset is shown in Figure 2a).

The time-to-event variable is the elapsed time calculated by the difference between the policy issue date and the event date. In case of right censored, this variable is calculated by the difference between the policy issue and the study closing date.

The principal statistics of the explanatory and response variables are indicated in Table 1 (categorical variables) and 2 (interval type variables).

Figure 2a) indicates there is a higher proportion of policyholders that have lapsed before their expiration.

Figure 2b) shows the distribution of the response variable time-to-event. There are higher rates of lapses at the beginning of the insurance contract. A peak is reached early on, usually within the first two years, after which the lapse rates decreases rapidly (Bauer, Gao, Moenig, Ulm, and Zhu, 2017).

A correlation analysis was performed as follows, Figure 3 summarizes all these correlations in differentiated groups.

- Using the Pearson coefficient to evaluate strength and direction of relationship among interval variables.
- For categorical and ordinal variables, the Cramer's V measure (Cramér, 1946) was determined.
- To conduct a correlation analysis between categorical and interval variables, the latter were transformed into binaries ones (i.e.,1 if the value is over the median, 0 otherwise) and thereafter, the Cramer's V correlation was determined.

From the original dataset, the initial insured capital and final insured capital variables were excluded because they yielded high correlation between them and with the premium variables. In addition, the option benefit variable has a perfect correlation with the product type and was removed from the study. The remaining

correlation coefficients among variables do not yield high values and are included in the study. There are not *missing* values in the dataset. Categorical variables are transformed into dummy (1,0) according to the number of categories indicated in Table 1. Payment frequency is the only ordinal variable in the dataset.

Table 2.
List of interval variables in our dataset

Variable	Statistics	Comments
Issue date	from 15/07/2005 to 27/12/2019	For policies not terminated by 31/12/2019 apply right censored
Time	Min 0, Mean 960.48, Max 5,249, std 960	In days basis
Age	Min 18, Mean 38.18, Max 79	
Annual premium	Mean 219.32, std 646.53	In American dollars
Supplementary premium	Mean 40.20, std 522.50	For other supplementary coverages (disability, health, etc.)

To apply the models RSF, CIF and CPH (as explained above), the data was randomly divided into training (70%) and test (30%) datasets.

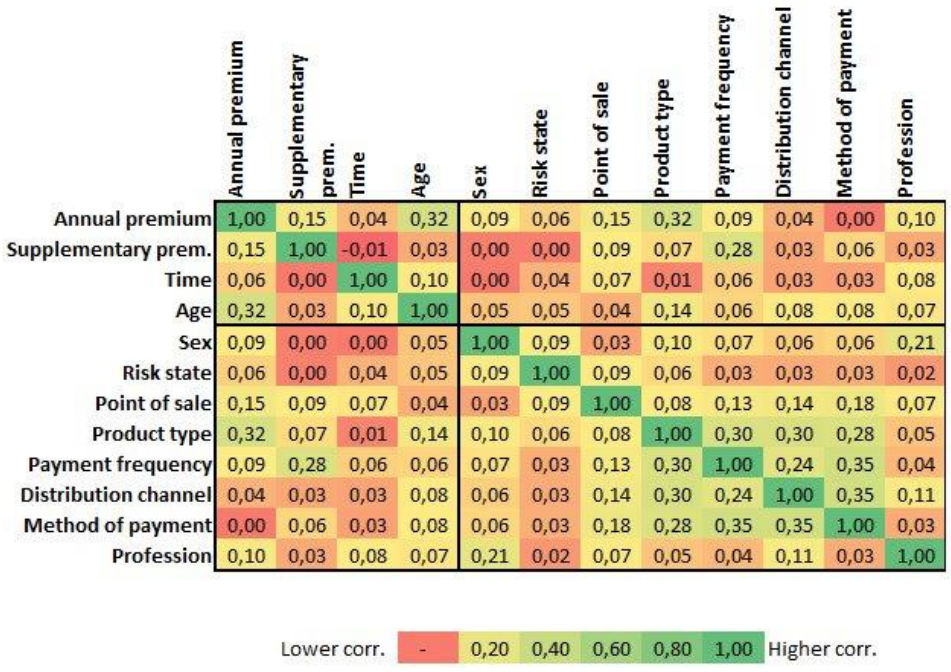

Figure 3. Correlation matrix: Cramer's V and Pearson (top left) coefficients according to the type of the variable

4. Results and analysis

This section presents and discusses the results of the RSF and CPH models. We focus on the consistency of RSF results (where variable importance and partial dependence of the covariates are computed), as compared to the traditional Cox model estimation. Additionally, for practical purposes, direct application using RSF model are shown such as estimating individual survival functions and creating risk groups.

The next section demonstrates that the performance of the machine learning techniques is better than the traditional CPH model for this dataset. Given that the results of the CIF and RSF techniques are similar, the focus throughout will be on the RSF model results alone.

4.1 Statistical analysis

The **Random Survival Forests** results and settings are summarized in Table 3. Note that *n-split* parameter has been set to a small number to avoid bias to the continuous variables in the variable importance calculation (Ishwaran and Kogalur, 2017).

Table 3
RSF: Results and settings

Training dataset size:	27,700
Number of lapses:	17,297
Number of trees:	1
Forest terminal node size:	1,000
Average no. of terminal nodes:	353.75
No. of variables tried at each split:	4
Resample size used to grow trees:	27,700
Splitting rule:	Logrank
Number of random split points:	4
Prediction Error (PE):	39.11%

The importance of the explanatory variables or variable importance (VIMP) is calculated. For this, the values of the variable x^{th} are randomly permuted, thus losing the relationship with the response variable. For each variable x_j, VIMP is obtained as the difference between the *PE* of the permuted assignment and the

original ensemble. If the original variable has an important relationship with the response variable, the *PE* value increases proportionally. None of the variables exhibit negative effects, so all have been included in the estimation of the RSF model. The results are shown in Figure 4.

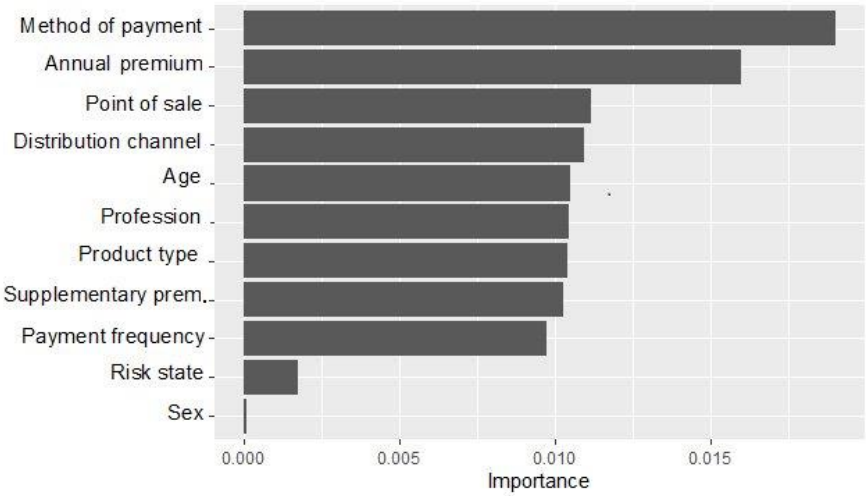

Figure 4. VIMP for explanatory variables using the permutation criterion

The partial dependence (marginal importance) of each explanatory variable has been obtained for the RSF model. For this, a value of the covariate x_j is fixed while adjusting for all other covariates then survival function $S_x(t \mid x)$ is calculated by equation (18). Figure 5 summarizes the 5-year predicted survival function and the marginal importance of each predictor variable over the estimated survival time.

$$S_x(t \mid x) = \frac{1}{n}\sum_{i=1}^{n} S_e(t \mid \mathbf{X}_i, x_{ji} = x) \tag{18}$$

Analysis of the influence that each of the explanatory variables has on the response variables is as follows:

- *Premium payment method*: Payment by bank debit reduces the probability of insurance lapse, and thus extends the permanence in the portfolio.
- *Premium amount*: Retention increases as amount of the premiums (both the main coverage and the supplementary ones) increases until these reach a peak then the retention decreases.

- *Point of sale*: There is a lower dispersion and a greater probability of insurance lapse in the city of Guayaquil than in the other cities, while in Quito, there is more dispersion and lower lapse probability.
- *Age*: The relationship noted between age and time spent in the portfolio confirms the emergency fund hypothesis that lapse rates are higher at younger ages.
- *Distribution channel*: The contribution of this variable to the survival function is similar for all underlying categories, even though the "Tied agent" category accounted for 80% of the sales in the portfolio. (Note that this variable was excluded in the Cox model, as it had no statistical significance).
- *Profession*: It's a reserved variable in our dataset, the contribution to retention is notably different across its categories.
- *Risk state*: Non-smoker insured present higher retention in the portfolio.
- *Sex*: This variable has a similar impact for both male and female categories in the RSF model, and in the CPH model it was excluded.

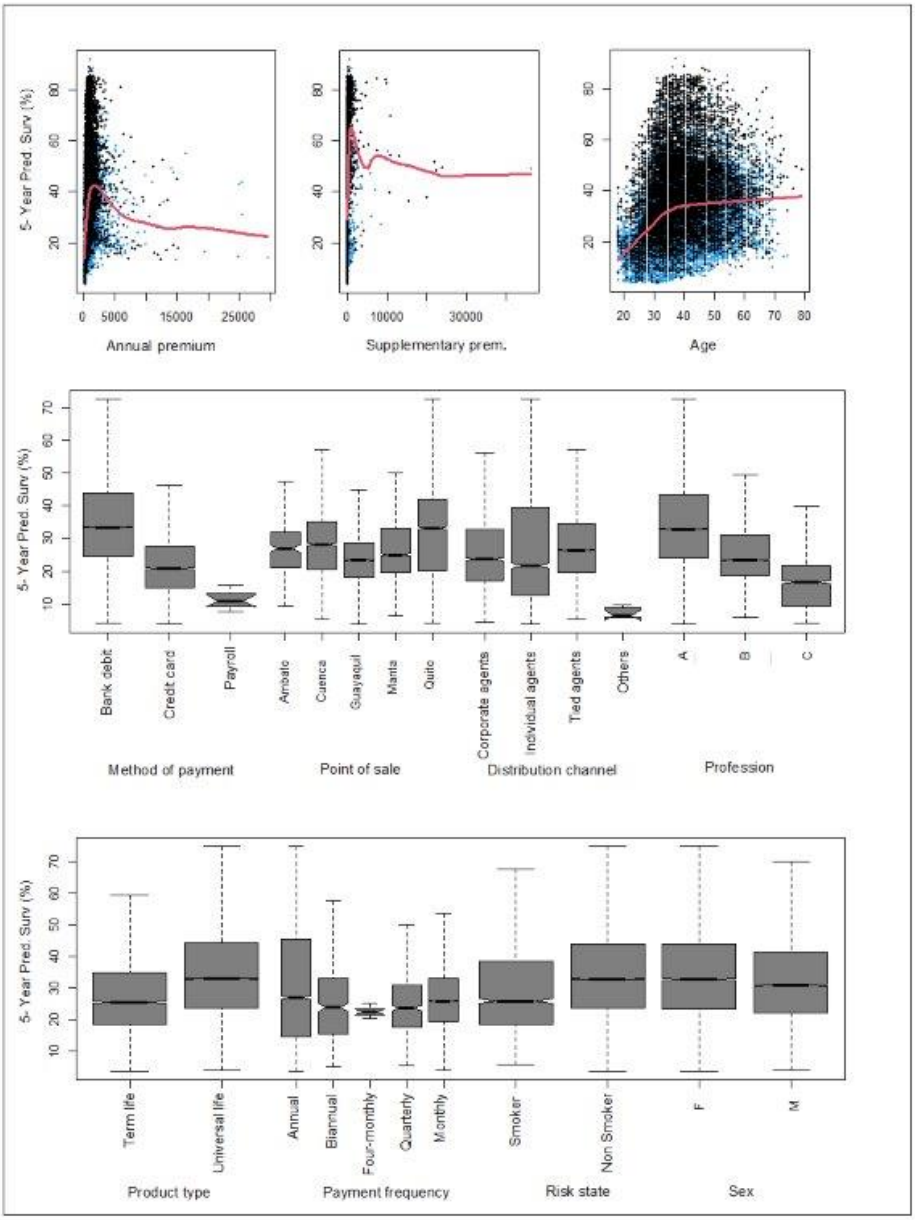

Figure 5. Marginal importance for explanatory variables: The Top Panels reflect the interval variables, and then categorical ones by VIMP importance

For practical actuarial purposes, the RSF model permits individual predictions of the CHF and survival function for each insured in the portfolio. An individual sample is dropped from the root in the training forest, and its survival estimators are calculated by the ensemble of the results at its terminal nodes. Refer to Figure 6a) for estimations of a random sample of policyholders in the portfolio.

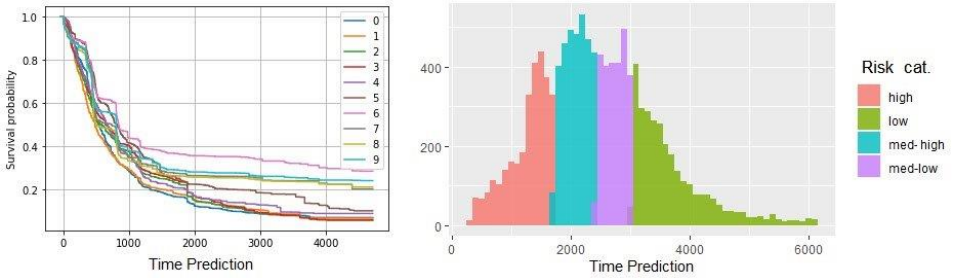

Figure 6. a) Predicted survival curves for ten randomly selected insured,and b) Segmentation of risk groups based on time predicted

Risk groups can be created to give a differentiated treatment in terms of risk management and create groups of homogeneous risk according to Solvency II, marketing campaigns, etc. In this case, four differentiated risk groups have been created according to the quartiles of the distribution of the estimated survival time of the test dataset, as shown in Figure 6b).

In the Cox Proportional Hazards model, all explanatory variables described in Tables 1 and 2 are considered. Regarding "dummyfication" process previously explained, a reference level is chosen for each of the categorical variables, and the explanations and analysis are relative to each such reference level.

In a backward modeling exercise, the "individual agent" category in the "Distribution channel" covariate was excluded at a 10% level of significance, and the "Sex" covariate as well as the "Others" category in the "Distribution channel" covariate were similarly excluded at a 5% level of significance. Therefore, in the final CPH model, sex and distribution channel were not taken into account.

In Table 4 hazards rates and their *p*- values are shown. The analysis is relative to the level of reference of each variable.

For the *Product type* variable, the level of reference is the "Term life" product, which means the "Universal life" product has a 9.60% (1-hazard ratio) less risk of lapse than Term life assuming the remaining variables do not change.

For the *Frecuency of payment* variable, the risk of lapse increases relative to an increase in number of payments in a year. This means that at any point in time, there is a 7.03% higher probability of lapse for a policyholder paying monthly than for another paying quarterly.

The contribution to predicted survival time for each covariate is consistent with machine learning RSF model. The consistency is demonstrated because there is a perfect correspondence between marginal importance results in the RSF model and hazards ratios in the CPH model.

Table 4.
Cox Regression Results

Variables	Hazard ratio	z	Pr(z)	
Age	0,988	-13,399	e-16	***
Annual premium	1	-5,961	2.51e-09	***
Supplementary Prem.	1	-2,875	0.004037	**
Payment frequency	1,073	11,039	e-16	***
Product type Ref: Term life				
Universal life	0,904	-5,61	2,03e-08	***
Risk state Ref: Smoker				
Non smoker	0,819	-6,369	9,03e-12	***
Point of sale Ref: Guayaquil				
Ambato	0,741	-3,535	0.000407	***
Cuenca	0,815	-7,143	9.14e-13	***
Manta	0,908	-2,96	0.003080	**
Quito	0,729	-17,688	2e-16	***
Method of payment Ref: Payroll				
Bank debit	0,178	-5,718	1.08e-08	***
Credit card	0,301	-3,967	7.27e-05	***
Profession Ref: B				
A	0,815	-10,602	2e-16	***
C	1,446	9,383	2e-16	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '

4.2 Results assessment of predictive power

The results for the three proposed models were compared using two indicators ─the Concordance index (C-index) and the Brier Score─. Figure 7a) plots the comparative C-index at each point in time *t* in the test dataset for the RSF (61.3%), CIF (60.8%), and CPH (58.8) models, respectively.

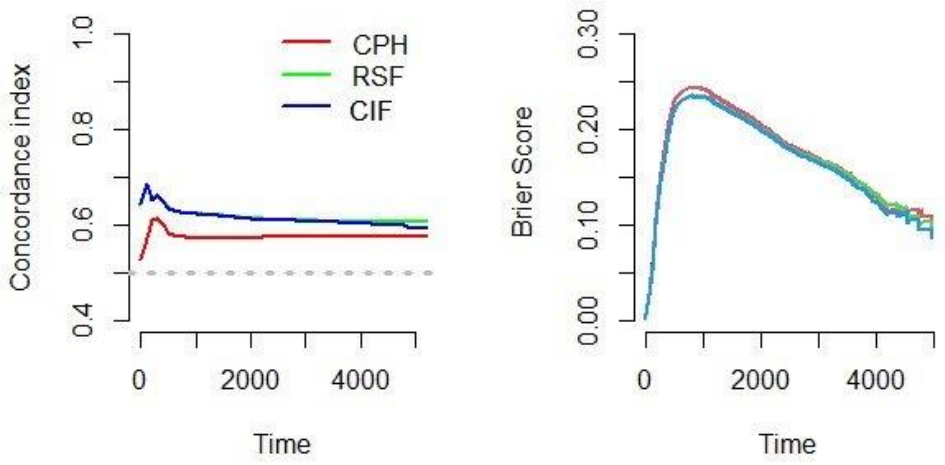

Figure 7. Comparative predictive power for three models: a) C-index, and b) Brier Score, at each point in time

Using the test dataset, the Brier Score is, 0.167 for RSF, 0.163 for CIF, and 0.170 for CPH. Figure 7b) plots the comparative results of the Brier Score at each point in time *t* for the three models.

Additionally, the three models and the comparative measures of predictive power ─namely, the C-index and Brier Score─ were implemented in PySurvival, an open source Python package for survival analysis modeling. The comparative results are similar because the base algorithm is identical (Wright and Ziegler, 2017).

5. Conclusions

This study implemented three models to estimate the survival function in an insurance portfolio, and assess the predictive power of each using the C-index and Brier Score measures, concluding that the RSF and CIF machine learning models perform better than the traditional CPH survival analysis model.

The results of the relationships between the explanatory variables and the time predicted by the survival function in the RSF and CPH models are consistent for all covariates. For example, in the "Method of payment" variable, the "Bank debit" category indicates a higher marginal importance than the other underlying categories, which is consistent with the lower hazard rate for this category in the CPH model. The "Sex" variable does not have a differentiated marginal effect across its underlying categories in the RSF model, and has the lowest value in the VIMP analysis; this is reflected in the CPH model, where the "Sex" variable does not have statistical significance.

The existing literature has been compared using the VIMP in the RSF model, among the main determinants for the exercise of the lapse option by a policyholder are the "premium amount" and "age" explanatory variables. In the dataset used in this study, among the main determinants of lapse are the "payment method", "point of sale", and "distribution channel" explanatory variables. On the other hand, "sex" and "risk state" variables are not important in the lapse determination.

For practical actuarial purposes, the RSF model allows for the estimation of individual probabilities of survival dissimilar at each future point in time, taking into account the insured explanatory variables; in this manner, the RSF better captures the heterogeneity of the characteristics of the insured. In addition, homogeneous risk groups may be also created for differentiated risk treatment, if required.

Prediction of survival function using the RSF model is widely used in biostatistical studies with good results, in this case its results are supported by CIF model, used the for first time to study time-to-event data in an insurance portfolio. The viability of the RSF and CIF machine learning techniques demonstrated in this study indicate that their application can be extended to the insurance industry.

The prediction of lapse rates using one model versus another could yield a significant variation in the determination of best estimate liabilities and solvency capital requirements for insurance companies, and therefore this is an important direction for future research.

Referencias

Aleandri, M., and Eletti, A. (2020). Modelling dynamic lapse with survival analysis and machine learning in CPI. *Decisions in Economics and Finance*, 1–20.

Bauer, D., Gao, J., Moenig, T., Ulm, E. R., and Zhu, N. (2017). Policy- holder exercise behavior in life insurance: The state of affairs. *North American Actuarial Journal*, *21*(4), 485–501.

Breiman, L. (2001). Random forests. *Machine learning*, *45* (1), 5–32.

Brockett, P., Golden, L., Guillén, M., Nielsen, J. P., Parner, J., and Perez-Marin, A. M. (2008). Survival analysis of a household portfolio of insurance policies: How much time do you have to stop total customer defection? *Journal of Risk and Insurance*, 75 (3), 713– 737.

Cramér, H. (1946). A contribution to the theory of statistical estimation. *Scandinavian Actuarial Journal*, *1946*(1), 85–94.

Crumer, A. M. (2011). Comparison between Weibull and Cox proportional hazards models. *Kansas State University Report*, 1–34.

EIOPA. (2010). EIOPA Report on the Fifth Quantitative Impact Study (QIS5) for Solvency II.

Eling, M., and Kiesenbauer, D. (2014). What policy features determine life insurance lapse? an analysis of the german market. *Journal of Risk and Insurance*, 81(2), 241–269.

Eling, M., and Kochanski, M. (2013). Research on lapse in life insurance:What has been done and what needs to be done? *The Journal of Risk Finance*, *14*(4), 392–413.

Fotso, S. et al. (2019). PySurvival: Open source package for survival analysis modeling. *Python package version 1.0*. https:// www. pysurvival.io/

Gerds, T. (2017). Pec: Prediction error curves for risk prediction modelsin survival analysis. *R package version*, *2*(4).

Hothorn, T., Hornik, K., and Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics*, *15*(3), 651–674.

International Accounting Standards Board. (2017). IFRS17: Insurance contracts.

Ishwaran, H., and Kogalur, U. (2017). Randomforestsrc: Random forests for survival, regression and classification (rf-src). 2016. *R package version*, *2*(0).

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., and Lauer, M. S. (2008).Random survival forests. *The annals of applied statistics*, *2* (3), 841–860.

Lee, E. T., and Wang, J. (2013). *Statistical methods for survival data analysis* (Vol. 476). John Wiley and Sons.

Mogensen, U. B., Ishwaran, H., and Gerds, T. A. (2012). Evaluating random forests for survival analysis using prediction error curves. *Journal of statistical software*, *50*(11), 1.

Pinquet, J., Guillén, M. and Ayuso, M. (2011). Commitment and lapse behavior in long-term insurance: A case study. *Journal of Risk and Insurance*, *78*(4), 983–1002.

Therneau, T. M., and Lumley, T. (2015). Package 'survival'. *R Top Doc*, *128*(10), 28–33.

Wright, M. N., and Ziegler, A. (2017). Ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, *77*(1), 1–17.

Zeileis, A., Hothorn, T., and Hornik, K. (2010). Party with the mob: Model-based recursive partitioning in R. *R package version 0.9*.