APPLICATIONS OF METAHEURISTICS IN INSURANCE

LÁSZLÓ KOVÁCS

Department of Computer Science, Corvinus University of Budapest, Hungary
E-mail: laszlo.kovacs2@uni-corvinus.hu

When calculating different profitability measures for a life insurance company, one of the most important parameters to know is the probability of a policy being in force at any given time after the start of risk bearing. These probabilities are given by the survival function. In this paper, we examine data from a Hungarian insurance company, in order to build models for the survival functions of two life insurance products. For survival function estimation based on the unique parameters of a new policy, Cox regression is used. However, not all parameters of a new policy are relevant in estimating the survival function. Therefore, application of model selection algorithms is needed. Furthermore, if the exact effects of the policy parameters for the survival function can be determined, the insurance company can direct its sales team to acquire policies with positive technical results. When traditional model selection techniques proposed by the literature (such as best subset, stepwise and regularization methods) are applied on our data, we find that the effect of the selected predictors for survival cannot be determined, as there is a harmful degree of multicollinearity. In order to tackle this problem, we propose adding the hybrid metaheuristic from Láng et al. (2017) to the Cox regression in order to eliminate multicollinearity from the final model. On the test sets, performance of the models from the metaheuristic rivals those of the traditional algorithms with the use of noticeably less predictors. These predictors are not significantly correlated and are significant for survival, as well. It is shown in the paper that with the application of metaheuristics, we could produce a model with good predicting capabilities and interpretable predictor effects. These predictor effects can be used to direct the sales activities of the insurance company.

Keywords: metaheuristics, Cox regression, insurance, model selection, regularization methods, best subsets

JEL-codes: G22, C34, C52, C63, C65

1. INTRODUCTION

When calculating different profitability measures for a life insurance company, one of the most important parameters to know is the probability of a policy being in force at any given time after the start of risk bearing. This probability is called the survival function of the policy. During the pricing of insurance products, actuaries make assumptions for the average survival function of the product, but these assumptions do not meet reality. From this difference, the company has a so-called technical result, which should be positive. The assumptions are mainly made for mortality and lapses. In order to calculate this result, we need to estimate the survival function for each policy and then match it with the assumed average. For survival function estimation based on the unique parameters of a new policy, Cox regression is used (Grosen – Jørgensen 2000). However, not all parameters of a new policy are relevant in estimating the survival function. So, application of model selection algorithms is needed in order to determine the relevant parameters of our new policy for estimating the survival function accurately. Furthermore, if the exact effects of the policy parameters for the survival function can be determined, the insurance company can direct its sales team to acquire policies with positive technical results.

In this paper, we examine data from a Hungarian insurance company, in order to build models for the survival functions of two life insurance products. In the dataset there are 39 policy parameters that can be used as predictors for new policies' survival functions. Some of these parameters are highly correlated, e.g. the existence of additional health coverages and its premiums or the sales channel, the change in acquiring and maintaining agents, and so on. When traditional model selection techniques proposed by the literature, such as best subset algorithms (Calcagno – de Mazancourt 2010; Furnival – Wilson 1974; Láng – Kovács 2014; Minerva – Paterlini 2010), stepwise (Lumley – Therneau 2004; Zhang 2016) and regularization (Allen 2013; Breheny 2013; Fan – Li 2002; Fan et al. 2010; Leng – Zhang 2006; Saldana – Feng 2018; Tibshirani 1997) methods are applied on our training data, we can find that they have good predictive results for the survival curves on a test set. On the other hand, the effect of the selected predictors for survival cannot be determined, as there is a harmful degree of multicollinearity and a huge number of insignificant predictors in our final models. For example, if we cross reference the different models, we cannot determine what additional coverages have positive or negative effects on survival. If we get some consistent results from our models, then they are not always reliable as they contradict pairwise statistical tests or reason. These features of the received models mean that we cannot direct the sales team in the kinds of policies they should strive to obtain. In order to tackle this problem, we adapted the hybrid metaheuristic

approach from Láng et al. (2017) to the Cox regression in order to eliminate multicollinearity from the final model. On the test sets, performance of the models from the metaheuristic rivals those of the traditional algorithms. However, the hybrid algorithm uses noticeably less predictors to achieve this. These predictors are not significantly correlated and are significant for survival, as well. All of this means that with the application of metaheuristics, we could produce a model for our data with good predicting capabilities and interpretable predictor effects. These predictor effects can be used to direct the sales activities for the company, e.g. we can suggest which additional coverages will produce better survival and through it, better technical results can be expected.

The paper is structured as follows. In Section 2, we introduce the basics of the Cox proportional hazard model. In Section 3, we give a quick overview of the model selection algorithms which are examined in the paper. Section 4 introduces the database used for the testing of the algorithms described in Section 3. Sections 5 and 6 contain the performance results for the different model selection methods. In Section 6, the advantages of the hybrid algorithm are also shown in the context of the interpretability for the survival models. In Section 7, the results are summarized with a focus on applications in directing insurance sales activity. The directions for further development are also drawn up.

2. INTRODUCTION TO THE PROBLEM

In order to calculate different profitability measures for a life insurance company, the most important parameter to determine is the number of policies at risk (marked r_x), which gives us the percentage of in-force policies after a given x time from the start of risk bearing. In-force policies can cease because of two reasons: death of the insured or the lapse/maturity of policy (Sheldon – Smith 2004).

During the premium calculation, mortality is considered through yearly life tables published by statistical offices and through the claims experience of the company, while the number of lapses is determined through assumptions or past experience. However, mortality and lapses always differ from assumptions made when calculating the premium. From this difference comes one of the most important sources of profit for insurers, the technical result. Because of this, it is important to estimate r_x in a way that a positive technical result is achieved (Grosen – Jørgensen 2000). For this reason, we aim to estimate r_x for an exact policy as accurately as possible. Furthermore, it is a serious challenge when creating a sales plan to determine the scope and effect of variables that can influence r_x , in order to specify what kind of new policies will be profitable. If we can describe what

kind of policies are preferred with variables known at the start of risk bearing, we can incite the sales team to acquire profitable policies.

The mortality component of r_x can be more precisely estimated by the inclusion of the cohort effect, for example with the use of Renshaw–Haberman model. However, our article does not investigate this problem.

Our main objective is to estimate the number of lapses based on different risk factors (regressors or independent variables) known at the time of acquisition. More precisely, we aim to estimate the percentage of not lapsed policies after a given x time from the start of risk bearing. The best tool to solve this problem is the application of Cox regression or Cox proportional hazards model (Grosen – Jørgensen 2000).

In Cox regression, T denotes time passed until a specified event (in our case lapse) as a random variable with a $G(t) = P(T \geq t)$ survival function. $G(t)$ gives us the probability of the event that our specified event (lapse) has not happened yet at time t . It is easy to see that $G(t)$ and r_x have a similar meaning. The only difference between the two functions is that in our interpretation r_x includes more events, not just lapses. In a Cox proportional hazard model, we can determine how the marginal change in a regressor can parallelly displace $\ln(h(t))$. Where $h(t) = \lim_{\varepsilon \rightarrow 0+} \frac{P(T < t + \varepsilon | T \geq t)}{\varepsilon}$, there is the probability that the examined event occurs right after time t , on the condition that the event has not occurred until

t . It can also be shown that $h(t) = \frac{(1 - G(t))'}{G(t)}$. So, the effect of the regressors is constant in time, they can only displace $\ln(h(t))$ parallelly. From this property comes the name, proportional hazards model (Cox 1972). Naturally, if we consider the interaction of a regressor with t , this assumption of proportionality can be resolved (Gillespie 2006).

With the help of the Cox regression, we can estimate the lapses component of r_x with the presence of different regressors. Based on the Cox model, we can determine the effect of the regressors on the behavior of lapses, if the regressors are significant and uncorrelated. The main task ahead is the selection of appropriate variables as regressors.

3. DETAILS OF THE APPLIED METHODOLOGY

In the Cox regression model, we aim to estimate $\ln(h(t))$ as the linear combination of our p possible regressors: $h(t | \mathbf{x}) = h_0(t) \exp(\boldsymbol{\beta}^T \mathbf{x})$ where $\mathbf{x} = (x_1, \dots, x_p)$ is the vector of our regressors and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)$ is the vector

of the coefficients to be estimated in the model. $h_0(t)$ is the baseline hazard rate which is our estimated hazard rate when the value of all our regressors is 0.

In survival analysis, we usually have a sample the size of n in the following structure: $(t_1, \mathbf{x}^{(1)}, d_1) \dots (t_n, \mathbf{x}^{(n)}, d_n)$ where t_i is the time the i th element was under our observation, and $\mathbf{x}^{(i)} = (x_{i1}, \dots, x_{ip})$ is the vector of the possible regressors for the i th element in our sample, d_i is the state variable, which equals to 1 if the examined event occurred in t_i time and equals 0 if the event has not happened until t_i , so the survival time until the event is censored in the sample.

If a sample in the previously given structure is available to us, then we can estimate β via maximizing the partial log likelihood function of our sample:

$$\ell(\beta) = \sum_{i: d_i=1} (\beta^T \mathbf{x}^{(i)} - \log \sum_{j: t_j > t_i} \exp(\beta^T \mathbf{x}^{(j)})) \quad (\text{Cox 1975}).$$

Once β is estimated, our task is to select a $\tilde{\mathbf{x}} \subseteq \mathbf{x}$ subset with $m \leq p$ elements in a way that the corresponding Cox regression model is parsimonious. A parsimonious model means that it can maximize the partial likelihood while using as few regressors as possible. The Akaike Information Criterion $AIC = -2\ell(\beta) + 2m$ and the Bayesian-Schwarz Information Criterion $SBC = BIC = -2\ell(\beta) + 2m \ln(\sqrt{n})$ are measures for the parsimony of a model generated by a given $\tilde{\mathbf{x}} \subseteq \mathbf{x}$ set. As we can see, these criteria are to be minimized as this way they reward the increase in the partial likelihood, while simultaneously penalizing the increase in the number of regressors used (m).

There are several algorithms in the literature that aim to select regressors from a given $\mathbf{x}$ set in order to achieve a parsimonious model with $\tilde{\mathbf{x}}$.

The most simple algorithm group is the stepwise-family. The two main members in this group are backward and forward selection algorithms.

The backward method initializes with all p variables used as regressors, then one by one eliminates a regressor, the omission of which decreases the used aim function (AIC or BIC). The algorithm terminates when the omission of a regressor cannot decrease the aim function further.

The forward method works the opposite way: it starts with no regressors in the model, and starts adding variables as regressors one by one from a given $\mathbf{x}$ set. In each iteration, the variable which can decrease the model's aim function the most is added as a regressor to $\tilde{\mathbf{x}}$. The algorithm terminates if inclusion of another variable can no longer decrease the chosen aim function.

These algorithms are simple and quick even in high dimensions (n and p). However, their main setback is that by giving or taking only one variable at a time, they can only explore a small portion of the search space, becoming local search algorithms. Because of this property of the stepwise methods, they are unlikely to find a global optimum for $\tilde{\mathbf{x}}$.

Of course, it is possible to develop an algorithm which tries to shift through every possible subset of $\mathbf{x}$ and find the subset with minimal *AIC* or *BIC*.

The first algorithm to attempt this is the leaps and bounds algorithm by Furnival and Wilson (1974), which applied the branch and bound approach, used to solve Mixed Integer and Linear Programming tasks, to the variable selection problem. Because examining all the possible subsets is a numerically challenging task ($2^p - 1$ different subsets to examine), the leaps and bounds algorithm usually does not find satisfying solutions for problems where $p > 20$.

The best subset selection algorithms applicable in the case of $p > 20$ are based on metaheuristics. One of the most popular metaheuristics to use is the genetic algorithm (Genetic Algorithm for Regressors' Selection, GARS for short, see Minerva – Paterlini 2010). Its main idea is to generate a random population from the possible subsets and starting from this initial population, we deduce new ones by copying the fittest elements (those with the best aim function values) to the new population and filling the remaining places by crossing these fittest solutions over and mutating the solutions in this new population by changing the inclusion property of some variables in the solution with a predefined probability. The algorithm terminates when the best solution in the current population has not changed for a given number of iterations.

The approach of GARS can ensure the finding of a solution that is close to the optimal one in a reasonable time. However, because the crossover operation of the GARS implicitly assumes a logical order of the possible regressors in the database (variables that are included as regressors are concentrated in certain columns of the data matrix), the algorithm can become slower than ideal. In these cases, an improved harmony search based solution (IHSRS for short, Láng – Kovács 2014) is proposed, where the crossover operator is replaced with the “playing from memory” operator, that randomly selects a good enough solution from the previous population instead of crossing over two, and mutates this solution more diversely than the mutation operator of the GARS. Then, if the new solution created this way has a preferable aim function value to the worst solution in the population, then the worst solution is replaced with the newly created one. It is also possible with a given probability that a completely new random solution is generated instead of mutating an existing one. With this it eliminates the assumption of concentrating variables in the data matrix and speeds up the algorithm.

Another approach to variable selection is the regularization family of methods. The most general case of these algorithms is the cosso (Leng – Zhang 2006), which assumes a non-linear Cox model: $h(t|\mathbf{x}^{(i)}) = h_0(t) \exp(\eta(\mathbf{x}^{(i)}))$, where $\eta(\cdot) \in \mathcal{F}$ and $\mathcal{F}$ is Reproducing Kernel Hilbert Space (RKHS for short). In this setting the partial log likelihood function to maximize for the estimation of the

coefficients takes the following form: $\ell = \sum_{i:d_i=1} (\eta(\mathbf{x}^{(i)}) - \log \sum_{j:t_j \geq t_i} \exp(\eta(\mathbf{x}^{(j)})))$.

In case of the cosso method, the coefficients of the model are estimated by minimizing the opposite of a penalized partial log likelihood function:

$$\min_{\eta \in \mathcal{F}} - \sum_{i:d_i=1} (\eta(\mathbf{x}^{(i)}) - \log \sum_{j:t_j \geq t_i} \exp(\eta(\mathbf{x}^{(j)}))) + \lambda \sum_{\alpha=1}^p \|P^\alpha \eta\|, \text{ where } P^\alpha \eta \text{ is the projec-}$$

tion of η to the orthogonal subspace of $\mathcal{F}$ generated by the α th possible regressor: $\mathcal{F} = \{1\} \oplus_{\alpha=1}^p \mathcal{F}_\alpha$, and $\|\cdot\|$ is the L_1 norm of $\mathcal{F}$. $\lambda \in (0, 1)$ is a parameter to be set. By estimating the coefficients of the model through a partial log likelihood penalized by the L_1 norm of the model parameters, it can be achieved that some $P^\alpha \eta$ projections are constant zero which, thus, executes variable selection.

In the linear case, when $\eta(\mathbf{x}) = \beta_0 + \sum_{i=1}^p \beta_i x_i$, then $\sum_{\alpha=1}^p \|P^\alpha \eta\|$ simply becomes $\sum_{i=1}^p |\beta_i|$, and the method is called the lasso method (Tibshirani 1997).

If $\eta(\mathbf{x})$ is linear and the $P^\alpha \eta$ projections are in a form of $p_\lambda(|\beta_j|)$, where $p_\lambda(\cdot)$ is a function of class SCAD (defined in Fan – Li 2002), then we get the SCAD method for a variable selection.

Although the regularization methods of regressor selection are considered to be one of the most novel methods for the task (Hastie et al. 2011), it can be shown that they also tend to underperform in the case of large datasets ($p > 20$) (Fan et al. 2010). For this reason, the so called iterative sure independence screening algorithm (ISIS for short) was developed for regressor selection (Fan et al. 2010).

The main idea of the ISIS algorithm is to assign a utility function to each possible regressor in $\mathbf{x}$ based on the information the variable contains on survival. The best d elements from $\mathbf{x}$ are selected and SCAD selection is applied on them. The selected variables make up the $\mathbf{x}'$ vector. For the variables that are not selected by SCAD, a conditional utility is defined that shows how much additional information a variable contains on survival, on the condition that variables in $\mathbf{x}'$ are used as regressors in the Cox model. Based on this conditional utility, the best d elements are selected. On the union of these newly selected variables and $\mathbf{x}'$ SCAD model selection is applied. The selected variables by this “second SCAD” make the new $\mathbf{x}'$. Because of this, variables that were in the first $\mathbf{x}'$ can be left out by the “second SCAD”. Of the variables that are left out from the new $\mathbf{x}'$, the conditional utility is calculated again, and the algorithm starts to loop the steps described above. The algorithm terminates when the composition of $\mathbf{x}'$ does not change for a given number of iterations. Once the terminating criterion is satisfied, $\tilde{\mathbf{x}} = \mathbf{x}'$.

Because of the pre-selection of possible regressors with utility functions, the ISIS method is a feasible regularization based variable selection method on datasets with large dimensions.

The variable selection methods examined in this section so far have a common disadvantage: they are all sensitive to multicollinearity and/or can select a final regressor set that contains correlated variables. Because of this property of the models obtained from these variable selection algorithms, interpretation of regressor effects becomes quite problematic, as the sign of the regressor's coefficients are not necessarily appropriate in the case of multicollinearity. Furthermore, the set of omitted variables does not necessarily contain the most irrelevant ones. It is easily possible that a correlated variable with a true regressor takes the place of the true regressor in the model without an effect of its own on the survival time. In the case of some of the examined algorithms, there are exact proofs that the results of the algorithm are not reliable in the presence of multicollinearity (Zhao – Yu 2006; Jia – Ju 2010).

In order to solve this problem originating from multicollinearity, we propose the Hybrid Genetical – Harmony Search Algorithm in Láng et al. (2017), which can be applied to Cox regression as well.

The main idea in the hybrid algorithm is to minimize model error while ensuring that only significant and uncorrelated variables are present in the final model. The level of significance and harmful multicollinearity are set through α and the level of the *VIF* indicator for each variable by the user. First, we attempted to extend our IHSRS algorithm with these constraints, but we found that accounting for these conditions slows down significantly our improved harmony search based solution. For this reason, we developed a new algorithm, which is suitable for parallelization.

The main idea is to preserve the simultaneous generation of new individuals from the GARS method (copying the best solutions from the old population and creating new solutions for the remaining places by different operators), while replacing the crossover operator with the “playing from memory” operator and the “random solution generating” operator that are applied instead of each other with a given probability. This way, we still do not assume a sequential structure for the possible regressors in the data matrix (like the GARS method does) and can parallelize the new solution generation part of our algorithm, as we do not need to match a new solution to the worst in our population.

The flowchart of our hybrid algorithm is given in *Figure 1*.

The core part of the algorithm is the updating of the population (the Harmony Memory, HM for short). During the memory updating phase we generate the new individuals in three ways in every new iterations:

1. The best new individuals are selected from the old memory, roulette selection is applied, so we select the ones better than the average – average of only those where the *VIF* criterion and the significant criterion is satisfied. If there are

Figure 1. Flowchart for the Hybrid Genetical – Harmony Search Algorithm

Source: author

not enough such individuals, we simply choose the ones with better fitness function values than the global average.

This transmission rate was appropriate, because if we decreased it, then we found the good solutions in the early populations much more later. On the other hand, if we increased this rate, then we reached the stable convergence much later, as we got some solutions that did not satisfy the *VIF* criterion.

For the remaining places new individuals are selected in two ways:

2. With *HMCR* probability, a selected element from those chosen as better than average in step 1 is modified: with the mutation probability, the inclusion property of a variable is changed in the selected element. The mutation probability is decreasing in the function of the iteration numbers to 2–3%, because when we are closer the optimum value, we want to modify with less intensity. And the *HMCR* probability is increasing to 95–97%, because when we are closer to the optimum, we need less totally new elements.

3. With $1 - \text{HMCR}$ probability a totally new individual is generated and selected.

The algorithm is stopped when the last 100 best partial likelihood values of the population do not change – do not increase, and do not decrease. This convergence criterion is determined in this way as in this case we can consider a result as the global optimum.

In the following section, we use a practical problem from insurance to show that our hybrid algorithm can select regressors that are uncorrelated, significant at a given α level and the error of the Cox model defined by these regressors is in the range determined by the other regressor selection algorithms.

4. INTRODUCTION OF THE DATABASE

In this paper, the $G(t)$ survival function for lapses is estimated from the policies of two life insurance products in the portfolio of a Hungarian insurance company. The main difference between the two examined products is that Product A is a pure risk product, so it only pays the sum assured in the event of death. On the other hand, Product B pays back the highest paid annual premium times the term, in case the insured is alive at the end of the term.

Because of these different characteristics, we build two different models for the two products. It is important to note that overdue or unpaid premiums are the main reason for lapse in case of both products (*Table 1*). So, our main focus should be on selecting regressors that are able to determine the non-paying clients at the start of the policy term.

Table 1. Lapse Reasons for Products A and B

Lapse Reason (Product A)	Number	Proportions
Overdue Premium	10 680	85.00%
Unpaid Premium	258	2.10%
Surrender by Client	1 621	12.90%
Summary	12 559	100%

Lapse Reason (Product B)	No.	Proportions
Overdue Premium	1 625	85.40%
Unpaid Premium	100	5.30%
Surrender by Client	178	9.40%
Summary	1 903	100%

Source: author

After the necessary data cleansing measures, we have 20 247 policies to examine in the case of Product A. In the case of these policies, the lapse rate is 43.9% for non-payment and 6.63% for surrenders by client. At the same time in the case of Product B, we have a sample of 6 628 policies, where the lapse rate for non-payment is 25.83% and 2.63% for surrenders. So, we can easily observe that the possibility for a return of premium at the end of term increases the willingness to pay the insurance premiums regularly. This difference between the two products

Figure 2. Survival functions for the two examined products

Source: author

can be also observed in their average survival functions in *Figure 2*. In this estimate for $G(t)$, regressors are not yet used.

All in all, we have 39 possible regressors to use, the value of which are known at the start of the policy term (*Table 2*). Two of these variables (payment method=cheque and number of other policies at the company by client) are worth considering with time-interactions as well, based on our statistical tests at $\alpha = 5\%$. The breakdown of our time variable is discretized by monthly intervals. Maturity, death and any other termination of the contract that is not given as a lapse reason in *Table 1*, are treated as censoring events in our models. All policies are observed since their commencement. Yearly payment and contract terms given years are multiplied by 12 as our time variable is discretized by monthly intervals.

5. RESULTS I. THE PROBLEMS OF TRADITIONAL VARIABLE SELECTION ALGORITHMS

The algorithms presented in Section 3 are applied on a training set that is 80% of our data, selected randomly. The examined algorithms are tested on the remaining 20% of our initial data. Due to the heuristic nature of some of the algorithms examined, we ran each algorithm 30 times and present their best final model in *Tables 3* and *4*. The runtimes displayed here are the average runtimes for the 30 runs. The training and test data were resampled in each of the 30 runs.

For the evaluation of the selected models, Begg's C-index is used, which gives us the conditional probability for the time passed until lapse being actually short-

Table 2. The list of policy parameters as possible regressors

Parameters	Parameters
Term in Years	Accidental Hospital Coverage?
Business Discount	Accidental Operation Coverage?
Entry Age	Operation Coverage?
Sum Assured	Critical Illnes Coverage?
Region	Other Health Coverage?
Annual Premium	Transport Accidental Coverage?
Nationality of Client	Women Health Coverage?
Is Client Foreign?	Premium of Accidental Disability Coverage
Experience of Agent (in months)	Premium of Accidental Death Cov.
Sales Chanel	Premium of Accidental Hospital Cov.
Difference in Acquiring and Maintaining Agent	Premium of Accidental Operation Coverage
Difference in Acquiring and Maintaining Sales Chanel	Premimum of Operation Coverage
Payment Method	Premium of Critical Illnes Coverage
Payment Frequency	Premium of Other Health Coverage
Smoker Status	Premium of Transport Accidental Coverage
Health Category	Premium of Women Health Coverage
With Loan Contract?	Number of Other Life Products
Is Client a Company?	Number of Other Non-Life Products
Accidental Disability Coverage?	Number of Lapsed Life Products
Accidental Death Coverage?	Number of Lapsed Non-Life Products

Source: author

er for a policy that is predicted as shorter lived than another, based on our regres-
sors: $C = P(t_2 > t_1 \mid \eta(\mathbf{x}^{(1)}) > \eta(\mathbf{x}^{(2)}))$ (Begg et al 2000).

This discriminating accuracy, defined by the C-index, is between 0.7015 and 0.7338 in case of Product A, and for Product B, it is between 0.6654 and 0.6679 for the same set of examined selection algorithms. The detailed results are in *Appendices 1* and *2*.

All the calculations were executed in R, version 3.3.3.. The R functions used to implement the different model selection algorithms are given in brackets in *Appendices 1* and *2*. The calculations were run on the following computer configuration:

- Processor: Intel Core i7-4770k 3.50 GHz
- Memory: 16 GB DDR3 SDRAM

However, in the case of the final models from the examined algorithms, the selected regressors are often not significant and are correlated in all of the models.

Between our 39 possible regressors, there are many naturally correlated variable groups, such as: sum assured, premium, age or sales channel, difference in acquiring and maintaining agent/sales channel immediately after the start of the

policy's term or the existence of accidental disability, accidental hospital and accidental operation coverages or the existence of additional coverages and their premiums or the number of existing or lapsed other policies at the company.

All of these correlating variable groups are represented by more than one variable in all of the selection algorithm's final models, which makes the interpretation of these final models challenging. For example, we cannot tell clearly that the existence or premium of certain additional coverages increases or decreases the probability of lapse. This phenomenon makes clear communication with the sales team harder, as we cannot tell what kind of additional coverages' acquisition should be rewarded. Furthermore, it is also disturbing that in most of the models, sales channel is a significant regressor, but it would be more logical to say that a correlating variable, the difference in the acquiring and maintaining agent is more relevant for the variability in the probabilities of lapse. However, the inclusion of this variable as a regressor always comes together with the inclusion of the sales channel as well. Plus, most of the time the difference in the acquiring and maintaining agent is not a significant regressor, even if it is included together with the sales channel.

In the case of the non-correlating variables, we do not receive surprising results from the models of the examined algorithms. In case of Product A, the regressors that increase the risk for lapse the most are the following: payment by cheque (effect decreasing in time), smoking, number of other lapsed life products at the company. In case of Product B, we can also say that payment by cheque increases the probability of lapse, with a decreasing effect in time. The other risk factors for lapse of Product B are the following: semi-annual payment frequency, smoking, worse health category and number of other lapsed life products at the company.

Detailed tables for the common significant risk factors for the final models of the traditional algorithms are presented in *Tables 3 and 4*.

6. RESULTS II. HANDLING MULTICOLLINEARITY WITH THE HYBRID ALGORITHM

For the solution of the problems originating from multicollinearity, we apply the hybrid algorithm presented at the end of Section 3. As we have shown in Section 3, the hybrid algorithm can minimize model error while ensuring that only significant and uncorrelated variables are present in the final model. Convergence criterion and other parameters of the algorithm can also be set by the user before running the algorithm. In our application $\alpha = 5\%$ and $\forall VIF_j \leq 2$. The code for the hybrid algorithm's implementation in R is available through the following link: <http://tinyurl.hu/SFkl/>

Table 3. Common significant risk factors from the tarditional algorithms, Product A,

Increasing the Chance of Lapse	Decreasing the Chance of Lapse
Annual Premium	Entry Age
Sales Chanel = Company Network	Sum Assured
Payment Method = Cheque	Experience of Agent
Smoker Status = yes	Payment Method = Automatic Transfer
Acc. Disability Coverage = yes	Accidental Death Coverage = yes
Acc. Operation Coverage = yes	Transport Accidental Coverage = yes
Number of Lapsed Life Products	Premium of Accidental Disability Coverage
	Number of Other Non-Life Products

Source: author

Table 4. Common significant risk factors from the tarditional algorithms, Product B,

Increasing the Chance of Lapse	Decreasing the Chance of Lapse
Annual Premium	Payment Method = Automatic Transfer
Sales Chanel = Company Network	Accidental Death Coverage = yes
Difference of Acquiring and Maintaining Agent	Number of Other Life Products
Difference of Acquiring and Maintaining Sales Chanel	Number of Other Non-Life Products
Payment Method = Cheque	
Payment Frequency = Semi-Annually	
Smoker Status = yes	
Health Category = Standard	
Accidental Disability Coverage = yes	
Accidental Hospital Coverage = yes	
Premium of Accidental Death Coverage	
Premium of Other Health Coverage	
Number of Lapsed Life Products	

Source: author

The most frequent final models of the hybrid algorithm have a C-index that rivals those of the traditional models. For the model of Product A, we have a C-index of 0.7131 and the C-index in case of Product B is 0.6658. So, we are well within the range defined for the C-index by the traditional algorithms. The main result is that this discriminating accuracy is achieved with significantly less and uncorrelated regressors in one model.

The average runtime during the 30 runs for the hybrid algorithm is in the scale of that of the GARS algorithm: 1375.138 seconds for Product A and 541.852 seconds for Product B. This is achieved with minimal parallelization via the “doParallel” R package (Calaway – Wetston 2014). However, by using the Kreas package (Arnold 2017), better runtimes can be achieved in the future.

The detailed output for the two final models are presented in *Tables 5* and *6*.

The models of the hybrid algorithm confirm our earlier assumption that the difference in the acquiring and maintaining agent is a more relevant predictor for lapse than the correlated sales channel variable. So, the sales channel is not the variable that carries the information regarding lapses, but the fact whether there was a change of agent for the policy. Of course, this phenomenon can be more common in case of a given sales channel, however, the main indicator to monitor is the change in agent right after the acquisition of the policy.

We can also determine the optimal strategy for the acquisition of additional coverages as well from the models of the hybrid algorithm. In the case of Product A, the existence of accidental death and critical illness coverages decreases the risk of lapse, while the existence of accidental hospital coverage increases the risk of lapse. It is important to note that from the strongly correlated accidental disability, accidental hospital and accidental operation coverages variable group, the existence of accidental hospital coverage has the most significant effect on the risk of lapse on its own. However, in case of the traditional algorithms, with more

Table 5. Final model of the hybrid algorithm, Product A

	Coef	Exp(Coef)	Delta %	p-value
ENTRY_AGE	-7.03E-03	0.992995	-0.701%	0.00012
ANNUAL_PREMIUM	1.04E-06	1.000001	0.000104%	< 2e-16
AGENT_EXPERIENCE	-2.84E-03	0.997164	-0.284%	< 2e-16
LOAN_CONTRACT=Y	-1.52E-01	0.858988	-14.101%	0.00048
PAYMENT_METHOD=cheque	1.98E+00	7.242743	624.274%	< 2e-16
PAYMENT_METHOD=automatic transfer	-5.48E-01	0.578105	-42.190%	< 2e-16
DIFFERENCE_IN_INTERVEINING_AGENT	1.94E-01	1.214096	21.410%	3.40E-14
SMOKER_STATUS=Y	6.36E-01	1.88891	88.891%	< 2e-16
ACCIDENTTAL_DEATH_COVERAGE=1	-2.51E-01	0.778022	-22.198%	< 2e-16
ACCIDENTTAL_HOSPITAL_COVERAGE=1	3.22E-01	1.379885	37.988%	< 2e-16
CRITICAL_ILLNESS_COVERAGE=2	-8.45E-02	0.918972	-8.103%	0.00752
PREMIUM_ACCIDENTAL_OPERATION_COVERAGE	-5.09E-05	0.999949	-0.005%	< 2e-16
PREMIUM_HEALTH_COVERAGE	-4.65E-06	0.999995	0.000%	3.90E-06
PREMIUM_TRANSPORTATION_COVERAGE	-1.50E-05	0.999985	-0.001%	0.00423
OTHER_LIFE	3.93E-01	1.481418	48.142%	< 2e-16
LAPSED_NON_LIFE	1.72E-02	1.017349	1.735%	5.70E-05
time_cheque	-3.94E-02	0.961366	-3.863%	< 2e-16
time_OTHER_LIFE	-1.35E-02	0.986591	-1.341%	< 2e-16

Source: author.

Table 6. Final model of the hybrid algorithm, Product B, Source: Own Editing

	Coef	Exp(Coef)	Delta %	p-value
ANNUAL_PREMIUM	8.07E-07	1.000001	0.0001%	0.0007
SALES_CHANEL=OTHER	7.93E-01	2.210017	121.0017%	0.0307
SALES_CHANEL=OWN_NETWORK	4.09E-01	1.505312	50.5312%	4.80E-05
SALES_CHANEL=PREMIUM	7.44E-01	2.104336	110.4336%	0.0112
DIFFERENCE_IN_INTERVEINING_AGENT	6.08E-01	1.836754	83.6754%	2.20E-16
PAYMENT_METHOD=cheque	3.80E+00	44.70118	4370.1184%	< 2e-16
PAYMENT_METHOD=automatic transfer		0.188247	-81.1753%	< 2e-16
SMOKER_STATUS=Y	7.17E-01	2.048279	104.8279%	< 2e-16
ACCIDENTAL_DISABILITY_COVERAGE=1	3.27E-01	1.386801	38.6801%	4.40E-06
ACCIDENTAL_DEATH_COVERAGE=1	-5.97E-01	0.550461	-44.9539%	< 2e-16
ACCIDENTAL_DISABILITY_PREMIUM	-1.35E-05	0.999987	-0.0013%	0.0026
ACCIDENTAL_DEATH_PREMIUM	2.14E-05	1.000021	0.0021%	5.10E-05
OTHER_NON_LIFE	-3.56E-02	0.965026	-3.4974%	0.0041
time_cheque	-1.71E-01	0.842822	-15.7178%	< 2e-16

Source: author.

than one variable present from this group, the existence of accidental disability coverage always seemed more significant. But this disability benefit variable is also shown insignificant by pairwise log-rank tests, so this variable has no immediate effect on lapses, but an indirect effect through its correlation with the existence of accidental hospital coverage.

Furthermore, in the case of Product A, if we acquire accidental operation, health and transportation accidents coverages on a high enough premium, the risk of lapse can be decreased.

These two previous observations from the model of the hybrid algorithm can be proxies for the latent variable of income: if the client simply has some additional coverages, it is possible that they were only persuaded into these coverages by the agent and soon they will have an overdue premium problem because of the premiums of the additional coverages. But, if the client already agrees to a higher premium, they probably have a high enough income to pay it over time.

We also have significant regressors in the final models of the hybrid algorithm that are not present as significant regressors in all the final models of the traditional algorithms, but reason suggests that they are significant predictors for lapse. Based on the regressors of the final models from the hybrid algorithm, we can say that policies that come with a bank loan and policies acquired by a more

experienced agent and policies with an insured of higher entry age have a lapse rate that is lower than average.

It seems logical that policies with higher premiums for the base product generally have higher lapse rates, as these premiums are not only proportional with the sum assured, but they reflect the individuals own risk as well.

Finally, an interesting point of these models from the hybrid algorithm is that the number of previously lapsed life products is not a significant regressor. The number of other existing life products is a significant regressor instead that increases the probability of lapse with a decreasing effect in time. Furthermore, the number of previously lapsed non-life products at the company is a significant regressor as well in the models of the hybrid algorithm.

We assume that this phenomenon is present because of the strict restrictions in the algorithm for significance and multicollinearity. These restrictions might have caused the effect of the number of previously lapsed products to disappear, no matter how reasonable its effect on increasing the risk of lapse is. Furthermore, this variable is also present as significant in every model of the traditional algorithms. However, it is important to note that there is harmful multicollinearity between the number of lapsed and still existing policies at the company, so the results of the traditional algorithms are not completely reliable either.

In the case of Product B, we should note based on the model from the hybrid algorithm that there is no harmful multicollinearity between the existence and premiums of the accidental death and disability benefits. The existence of a coverage for accidental death decreases the probability of lapse, while an increase in its premium (more precisely: increase of its premium's ratio in the whole annual premium) increases the same probability. The directions for the effects of the existence and the premium of the accidental disability benefit are reversed.

Furthermore, in the case of strongly correlating variables for Product B, the hybrid model shows that mainly smoking habits influence lapses rather than the general health condition of a patient. It is also shown by the hybrid model that primarily the change in the acquiring and maintaining agent after the start of the policy is the more immediate risk for lapse rather than the correlating effect of change in the sales channel.

We can also observe that the increase of the annual premium for the whole policy is the main risk factor from the correlating variable group of age, sum assured and annual premium.

In the case of Product B, there is no significant effect on lapse from the non-accidental additional benefits on their own. We can approximate the discriminating power of the models from the traditional selection algorithms without using the existence or premiums of these accidental benefits as regressors.

Our final observation is that in the case of Product B, semi-annual payment frequency as a risk factor for lapse is not present in the final model of the hybrid algorithm, while its effect is always present in the final models of the traditional algorithms and it is not a member of any correlating variable group. Basically, the strict restrictions for significance and the discrete variable selection method of the hybrid algorithm lead to the omission of the whole payment frequency variable in the final model of the hybrid algorithm. So, we can identify a point in our algorithm that requires improvement in order to detect every significant effect for the target variable.

7. SUMMARY

This paper shows that one of the most important parameters for calculating profitability measures of a life insurance company is the number of policies at risk (r_x). In the paper, the estimation of r_x due to lapses is given for a policy with specific parameters (entry age, premium, history of client at the company, etc.) by a Cox regression in a sample from a Hungarian insurance company.

If we can determine r_x for each specific policy, we can determine a sales strategy that incites the acquisition of policies with small expected lapse rates, which usually have higher profitability. Of course, lapse rates are not the only factor that influences profitability. For example, high-premium policies may be encouraged even if they have somewhat higher lapse rates, because they may still lead to higher profit measures.

However, the traditional algorithms for model selection in a Cox regression are sensitive for multicollinearity, which means that regressors in the final models of these algorithms can be redundant and we cannot be sure that we selected the most relevant and logical variables as regressors from the correlating variable groups. Because of this phenomenon, we cannot clearly give sales agents advice on what kind of policies to acquire, despite having models with high discriminating power.

The problem is solved by the application of our Hybrid Genetic – Improved Harmony Search Algorithm for model selection in Cox regression. The discriminating power of the final model of the hybrid algorithm is comparable to those of the models from the traditional algorithms, while it uses considerably less and not redundant regressors to achieve this kind of discriminating power. This good performance is due to the algorithm's strict conditions that require significant and non-correlating variables in the final model.

In consequence, we can give exact parameters to the sales agents to look for while acquiring a new life insurance policy, based on the final model of the Hybrid algorithm.

REFERENCES

- Allen, G. I. (2013): Automatic Feature Selection via Weighted Kernels and Regularization. *Journal of Computational and Graphical Statistics* 22(2): 284–299.
- Arnold, T. B. (2017): KerasR: R Interface to the Keras Deep Learning Library. *The Journal of Open Source Software* 2.
- Begg, C. B. – Cramer, L. D. – Venkatraman, E. S. – Rosai, J. (2000): Comparing Tumour Staging and Grading Systems: A Case Study and a Review of the Issues, Using Thymoma as a Model. *Statistics in Medicine* 19(15): 1997–2014.
- Breheny, P. (2013): ncvreg: Regularization Paths for SCAD-and MCP-Penalized Regression Models. *R package* version 2.6-0.
- Calaway, R. – Weston, S. (2014): doParallel: Foreach Parallel Adaptor for the Parallel Package. *R package* version 1(8).
- Calcagno, V. – de Mazancourt, C. (2010): glmulti: An R Package for Easy Automated Model Selection with (Generalized) Linear Models. *Journal of Statistical Software* 34(12): 1–29.
- Cox, D. (1972): Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)* 34(2): 187–220.
- Cox, D. R. (1975): Partial Likelihood. *Biometrika*: 269–276.
- Fan, J. – Li, R. (2002): Variable Selection for Cox's Proportional Hazards Model and Frailty Model. *Annals of Statistics* 74–99.
- Fan, J. – Feng, Y. – Wu, Y. (2010): High-Dimensional Variable Selection for Cox's Proportional Hazards Model. In: Berger, J. O. – Cai, T. T. – Johnstone, I. M. (eds): *Borrowing Strength: Theory Powering Applications—A Festschrift for Lawrence D. Brown*. Beachwood, OH: Institute of Mathematical Statistics, pp. 70–86.
- Furnival, G. M. – Wilson, R. W. (1974): Regressions by Leaps and Bounds. *Technometrics* 16(4): 499–511.
- Gillespie, B. (2006): Checking Assumptions in the Cox Proportional Hazards Regression Model. *Presented at the Midwest SAS Users Group (MWSUG) Dearborn, Michigan*.
- Grosen, A. – Jørgensen, P. L. (2000): Fair Valuation of Life Insurance Liabilities: The Impact of Interest Rate Guarantees, Surrender Options, and Bonus Policies. *Insurance: Mathematics and Economics* 26(1): 37–57.
- Jia, J. – Yu, B. (2010): On Model Selection Consistency of the Elastic Net. *Statistica Sinica* 20: 595–611.
- Láng, B. – Kovács, L. (2014): Linear Regression Model Selection Using Improved Harmony Search Algorithm. *SEFBIS Journal* 9(1): 15–22.
- Láng, B. – Kovács, L. – Mohácsi, L. (2017): Linear Regression Model Selection Using a Hybrid Genetic – Improved Harmony Search Parallelized Algorithm. *SEFBIS Journal* 11(1): 2–9.
- Leng, C. – Zhang, H. (2006): Model Selection in Nonparametric Hazard Regression. *Nonparametric Statistics* 18(7–8): 417–429.
- Lumley, T. – Therneau, T. (2004): The Survival Package. *R News* 4(1): 26–28.

- Minerva, T. – Paterlini, S. (2010): Regression Model Selection Using Genetic Algorithms. *Proceedings of the 11th WSEAS International Conference on RECENT Advances in Neural Networks, Fuzzy Systems & Evolutionary Computing*: 19–27.
- Saldana, D. F. – Feng, Y. (2018): SIS: An R Package for Sure Independence Screening in Ultrahigh Dimensional Statistical Models. *Journal of Statistical Software* 83(2): 1–25.
- Sheldon, T. J. – Smith, A. D. (2004): Market Consistent Valuation of Life Assurance Business. *British Actuarial Journal* 10(3): 543–605.
- Tibshirani, R. (1997): The Lasso Method for Variable Selection in the Cox Model. *Statistics in Medicine* 16(4): 385–395.
- Vanderhoof, I. T. – Altman, E. (Eds.) (2013): *The Fair Value of Insurance Liabilities (Vol. 1)*. Springer Science & Business Media.
- Zhang, Z. (2016): Variable Selection with Stepwise and Best Subset Approaches. *Annals of Translational Medicine* 4(7): 136.
- Zhao, P. – Yu, B. (2006): On Model Selection Consistency of Lasso. *The Journal of Machine Learning Research* 7: 2541–2563.

Open Access. This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (<https://creativecommons.org/licenses/by/4.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited, a link to the CC License is provided, and changes – if any – are indicated. (SID_1)

APPENDIX 1 - Application of Traditional Model Selection Algorithms to Product A (R functions used are given in brackets)

Algorithm	Left out variables	Not significant variables, $\alpha=5\%$, $\alpha=10\%$	C-index	Correlated variables	Runtime (sec)
Full model (coxph)	---	TERM_YRS, BusDisc, REGIO=DelDunantul, REGIO=EszakAlfold, REGIO=EszakMagyar, REGIO=Kulfold, REGIO=NyugatDunantul, SALES_CHNL=PREMIUM, INTERVENIATING_DIFF=1, INTERV_SLSCHN_DIFF=1, PAYMENT_FREQ=semiann, PAYMENT_FREQ=quarterly, HEALTH_CATEGORY=S, CRIT_ILINESS=1, HEALTH_COV=1, WOMEN_HEALTH=1, ACC_DEATH_PREM, ACC_OPERATION_PREM, FOREIGN	0.73351	INIT_SA~PREMIUM INTERV_SLSCHN_DIFF ~ ACC_DISAB ~ ACC_HOSP, ACC_OPERATION COVER EXISTENCE ~ COVER PREM No of other policies ~ No of other lapsed policies	0.7577159
Stepwise (stepAIC)	BusDisc, INTERVENIATING_DIFF, INTERV_SLSCHN_DIFF, PAYMENT_FREQ, HEALTH_CATEGORY, HEALTH_COV, WOMEN_HEALTH, ACC_DEATH_PREM, FOREIGN	TERM_YRS, REGIO=DelDunantul, REGIO=EszakAlfold, REGIO=EszakMagyar, REGIO=Kulfold, REGIO=NyugatDunantul, SALES_CHNL=PREMIUM	0.73377	INIT_SA~PREMIUM ACC_DISAB ~ ACC_HOSP, ACC_OPERATION COVER EXISTENCE ~ COVER PREM No of other policies ~ No of other lapsed policies	200.44422
GARS (glmulti)	BusDisc, INTERVENIATING_DIFF, INTERV_SLSCHN_DIFF, PAYMENT_FREQ, HEALTH_CATEGORY, HEALTH_COV, WOMEN_HEALTH, ACC_DEATH_PREM, FOREIGN	TERM_YRS, REGIO=DelDunantul, REGIO=EszakAlfold, REGIO=EszakMagyar, REGIO=Kulfold, REGIO=NyugatDunantul, SALES_CHNL=PREMIUM	0.73377	INIT_SA~PREMIUM ACC_DISAB ~ ACC_HOSP, ACC_OPERATION COVER EXISTENCE ~ COVER PREM No of other policies ~ No of other lapsed policies	1393.0524
Lasso (ncvsurv)	REGIO=NyugatDunantul, PAYMENT_FREQ=semiann, WOMEN_HEALTH, ACC_DEATH_PREM	not defined	0.73330	INTERVENIATING_DIFF ~ INTERV_SLSCHN_DIFF ACC_DISAB ~ ACC_HOSP, ACC_OPERATION COVER EXISTENCE ~ COVER PREM No of other policies ~ No of other lapsed policies	6.339347

Algorithm	Left out variables	Not significant variables, $\alpha=5\%$, $\alpha=10\%$	C-index	Correlated variables	Runtime (sec)
SCAD (ncvsurv)	TERM_YRS, BusDisc, INIT_SA, REGIO, SALES, CHN=PREMIUM, INTERV_ SLSCHN_DIFF, PAYMENT_ FREQ, HEALTH_CATEGORY, LOANCOVER, OPERATION_ COVER, CRIT_ILINESS, COVER, WOMEN_OPERATION, WOMEN_HEALTH, ACC_OPERATION_PREM, HEALTH_COVER_PREM, TRANSPORT_PREM, WOMEN, HEALTH_PREM, OTHER_LIFE, LAPSED_NON_LIFE, FOREIGN	not defined	0.70150	INIT_SA~PREMIUM ACC, DISAB ~ACC_OPERATION ACC_DISAB EXISTENCE ~ ACC_DISAB PREM	22.29408
ISIS (SIS)	REGIO=EszakAlfold, REGIO=NyugatDunantul, PAYMENT_FREQ=semiann, WOMEN_HEALTH, ACC_ DEATH_PREM	not defined	0.73166	INIT_SA~PREMIUM INTERVENIATING_DIFF ~ INTERV_SLSCHN_DIFF ACC_DISAB ~ACC_HOSP, ACC_OPERATION COVER EXISTENCE ~ COVER PREM No of other policies ~ No of other lapsed policies	14.06437
Cosso (cosso)	REGIO=NyugatDunantul, PAYMENT_FREQ=semiann, WOMEN_HEALTH, ACC_ DEATH_PREM	not defined	0.73385	INIT_SA~PREMIUM INTERVENIATING_DIFF ~ INTERV_SLSCHN_DIFF ACC_DISAB ~ACC_HOSP, ACC_OPERATION COVER EXISTENCE ~ COVER PREM No of other policies ~ No of other lapsed policies	1957.955

Source: authors

APPENDIX 2 - Application of Traditional Model Selection Algorithms to Product B (R functions used are given in brackets)

Algorithm	Left out variables	Not significant variables, $\alpha=5\%$, $\alpha=10\%$	C-index	Correlated variables	Runtime (sec)
Full model (coxph)	---	TERM_YRS, BusDisc, ENTRY_AGE, INIT_SA, REGIO, AGENT_EXPER, SALES_CHN=OTHER, SALES_CHN=PREMIUM, INTERV_SLSCHN_DIFF, PAYMENT_FREQ=monthly, PAYMENT_FREQ=quarterly, LOANCOVER, ACC_OPERATION, OPERATION, CRIT_ILLNESS, HEALTH_COVER, WOMEN_HEALTH, ACC_HOSPITAL_PREM, ACC_OPERATION_PREM, OPERATION_PREM, CRIT_ILLNESS_PREM, HEALTH_COVER_PREM, TRANSPORTATION_COVER_PREM, WOMEN_HEALTH_PREM, FOREIGN	0.66728	ENTRY_AGE~INIT_SA~PREMIUM~SALES_CHN~INTERVENIATING~INTERV_SLSCHN_DIFF~HEALTH_CAT~SMOKER~ACC_DISAB~ACC_OPERATION+ACC_HOSPITAL~CRIT_ILLNESS~OPERATION+HEALTH_COVER~TRANSPORTATION_COVER~ACC_DISAB+ACC_DEATH~COVER_EXISTENCE~COVER_PREM~No of other policies~No of other lapsed policies	0.155159
Stepwise (stepAIC)	ENTRY_AGE, REGIO, AGENT_EXPER, LOANCOVER, ACC_OPERATION, OPERATION, CRIT_ILLNESS, HEALTH_COVER, TRANSPORTATION_COVER, WOMEN_HEALTH, ACC_OPERATION_PREM, OPERATION_PREM, CRIT_ILLNESS_PREM, TRANSPORTATION_PREM, WOMEN_HEALTH_PREM, FOREIGN	TERM_YRS, BusDisc, INIT_SA, SALES_CHN=OTHER, SALES_CHN=PREMIUM, INTERV_SLSCHN_DIFF, PAYMENT_FREQ=monthly, PAYMENT_FREQ=quarterly, ACC_HOSPITAL_PREM	0.667211	INIT_SA~PREMIUM~SALES_CHN~INTERVENIATING~INTERV_SLSCHN_DIFF~HEALTH_CAT~SMOKER~ACC_DISAB~ACC_HOSPITAL~ACC_COVER_EXISTENCE~ACC_COVER_PREM~No of other policies~No of other lapsed policies	78.25758

Algorithm	Left out variables	Not significant variables, $\alpha=5\%$, $\alpha=10\%$	C-index	Correlated variables	Runtime (sec)
GARS (glmulti)	TERM_YRS, INIT_SA, REGIO, AGENT_EXPER, LOANCOVER, ACC, OPERATION, OPERATION, CRIT_ILLNESS, HEALTH, COVER, TRANSPORTATION, COVER, WOMEN_HEALTH, ACC OPERATION_PREM, OPERATION_PREM, CRIT_ILLNESS_PREM, TRANSPORTATION_COVER, PREM, WOMEN_HEALTH, PREM, FOREIGN	SALES_CHANEL=OTHER, SALES_CHANEL=PREMIUM, PAYMENT_FREQ=monthly, PAYMENT_FREQ=quarterly, BusDisc, ENTRY_AGE	0.667526	ENTRY_AGE~PREMIUM SALES CHANEL~INTERVENIATING_ DIFF+INTERV SLSCHNL_DIFF HEALTH_CAT~SMOKER ACC DISAB~ACC_HOSPITAL No of other policies ~ No of other lapsed policies ACC COVER EXISTENCE ~ ACC_COVER PREMIUMS	417.342
	BusDisc, INIT_SA, REGIO=DelAlfold, REGIO=EszakMagyar, REGIO=KozepDunantul, REGIO=NyugatDunantul, SALES_CHANEL=OTHER, SALES_CHANEL=PREMIUM, PAYMENT_FREQ=monthly, PAYMENT_FREQ=quarterly, LOANCOVER, ACC, OPERATION, OPERATION, CRIT_ILLNESS, TRANSPORTATION_COVER, WOMEN_HEALTH, ACC DISAB_PREM, ACC_HOSPITAL_PREM, ACC_OPERATION_PREM, OPERATION_PREM, TRANSPORTATION_COVER, PREM, WOMEN_HEALTH, PREM, LAPSED_NON_LIFE, FOREIGN	not defined	0.66537	ENTRY_AGE~PREMIUM SALES CHANEL~INTERVENIATING_ DIFF+INTERV SLSCHNL_DIFF HEALTH_CAT~SMOKER ACC DISAB~ACC_HOSPITAL No of other policies ~ No of other lapsed policies ACC DEATH~HEALTH_ COVER EXISTENCE ~ ACC_ DEATH+HEALTH_COVER PREMIUMS	1.759689
Lasso (ncvsurv)					

Algorithm	Left out variables	Not significant variables, $\alpha=5\%$, $\alpha=10\%$	C-index	Correlated variables	Runtime (sec)
SCAD (ncvsurv)	INIT_SA, REGIO=DelAlfold, REGIO=EszakMagyar, REGIO=KozepDunantul, REGIO=NyugatDunantul, AGENT_EXPER, SALES_CHANEL=OTHER, SALES_CHANEL=PREMIUM, PAYMENT_FREQ=monthly, PAYMENT_FREQ=quarterly, LOANCOVER, ACC_OPERATION, OPERATION, CRIT_ILLNESS, TARNSPORTATION_COVER, WOMEN_HEALTH, ACC_HOSPITAL_PREM, ACC_OPERATION_PREM, OPERATION_PREM, CRIT_ILLNESS_PREM, TRANSPORTATION_COVERAGE_PREM, WOMEN_HEALTH_PREM, FOREIGN	not defined	0.66762	ENTRY_AGE~INIT_SA+PREMIUM SALES_CHANEL~INTERVENIATING_DIFF+INTERV_SLSCHN_DIFF HEALTH_CAT~SMOKER ACC_DISAB~ACC_HOSPITAL HEALTH_COVER+ACC_DISAB+ACC_DEATH_EXISTENCE~HEALTH_COVER+ACC_DISAB+ACC_DEATH_PREM No of other policies ~ No of other lapsed policies	9.395016
ISIS (SIS)	REGIO=DelDunantul, REGIO=EszakMagyar, REGIO=KozepDunantul, PAYMENT_FREQ=quarterly, OPERATION, CRIT_ILLNESS, WOMEN_HEALTH, ACC_OPERATION_PREM, WOMEN_HEALTH_PREM	not defined	0.66793	ENTRY_AGE~PREMIUM SALES_CHANEL~INTERVENIATING_DIFF+INTERV_SLSCHN_DIFF HEALTH_CAT~SMOKER ACC_DISAB~ACC_HOSPITAL+ACC_OPERATION TRANPOTATION_COVER~ACC_DISAB+ACC_DEATH No of other policies ~ No of other lapsed policies ACC_COVER_EXISTENCE~ACCCOVERPREMIUMS	4.10537

Source: author