MISCELLANEA

Bank i Kredyt 56(1), 2025, 113-138

BANK I KREDYT
BANK & CREDIT

Feature selection methods for Cox proportional hazards model. Comparative study for financial and medical survival data

Wojciech Skwirz*

Submitted: 4 May 2024. Accepted: 26 August 2024.

DOI: 10.5604/01.3001.0054.9612

Abstract

This study compares Cox proportional hazards models across medical and financial datasets built using various feature selection techniques. In this analysis 8 feature selection techniques (3 variants of forward selection, 2 variants of a selection based on principal component analysis, selection based on random survival forest, best subset selection and a selection based on a LASSO regularization) were tested across 22 multidimensional datasets (2 financial and 20 medical). The resulting Cox models were compared based on a concordance index. The main hypothesis of this study stating that the LASSO regularization or the selection based on random survival forest method (generating good models for medical data) would yield similar performance on financial data, was hereby disproved. The forward Schwarz and best subset selection gave the best results for financial data, while LASSO and random survival forest proved to be the most efficient in medical setting, for each considered model size.

Keywords: survival analysis, feature selection, credit risk, financial data, medical data

JEL: C24, C41, C52, C55, G21, I10

* Warsaw School of Economics; e-mail: wojciech.skwirz@gmail.com; ORCID: 0000-0002-6384-8143.

1. Introduction

Survival analysis is a group of statistical methods which are intended to estimate a time until an event of interest occurs. Historically, primary applications of survival modeling included domains such as industry, medicine, actuarial science or demographics.

Survival modeling has emerged as a cornerstone of the medical field, revolutionizing mortality studies and shaping the landscape of clinical trials for novel pharmaceutical interventions. By accounting for censoring and the influence of explanatory covariates, survival models enable researchers to uncover patterns and prognostic factors for the time of survival, thus providing invaluable insights into disease dynamics and treatment efficiency.

While survival modeling has found widespread application in fields such as medicine and demography, its implementation in the financial sector, particularly within banking, remains relatively limited.

In accordance with the International Financial Reporting Standard 9 (IFRS9)[1], a bank is required to create provisions for credit risk. The first step of this process consists in assigning stages to credit exposures. In particular, stage 2 classifies exposures with a significant increase in credit risk (SICR) since initial recognition. For these exposures, banks have an obligation to estimate the probability of default (i.e. a situation when credit obligations are not met) in a loan's lifetime horizon.

By employing survival modeling banks could build better models for credit risk provisions in accordance with IFRS9. Survival modeling holds promise for enhancing risk management practices and improving decision-making processes in financial institutions.

2. Main goal of the study

Real life datasets used in bioinformatics are often characterized by a very large number of explanatory variables with relatively few observation units. For example, while analyzing DNA microarrays for patients with a rare disease, often tens of thousands explanatory variables are available (n_m) with only hundreds of units (i_m). Meanwhile for financial data the situation is quite different. Banks gather vast amounts of characteristics of their customers which results in tens of thousands potential explanatory variables (n_f). As major financial institutions, banks also have a substantial client base which provides a large number of observation units – often counted in hundreds of thousands (i_f).

This difference can be summarized in the following points:

1) number of observation units: $i_f >> i_m$,

2) number of features: $n_f \approx n_m$,

3) both n_f and n_m can be in the range of tens of thousands.

This issue provides a range of possibilities for research aimed at improving the relevance of survival modeling for banks.

The main goal of this article is to determine which feature selection technique leads to the best (in terms of their predictive power) Cox proportional hazards models which could be used for credit risk assessment by banks.

[1] Commission Regulation (EU) 2016/2067 of 22 November 2016 amending Regulation (EC) No. 1126/2008 adopting certain international accounting standards in accordance with Regulation (EC) No. 1606/2002 of the European Parliament and of the Council as regards International Financial Reporting Standard 9 (text with EEA relevance).

Based on the literature overview (see section 3), there is no one universal recommended method for feature selection in survival analysis. However, the LASSO regularization and the selection based on random survival forest frequently yield good results in the case of Cox proportional hazards models used for medical research. The main hypothesis considered by this research assumes that these methods should also generate superior models when built on financial datasets. This hypothesis was verified by a comparative analysis of the performance of Cox proportional hazards models distinguished by various feature selection techniques. For this purpose specifically, the models were developed based on two sets of data from separate fields:

a) 2 datasets from the financial sector, with a large number of covariates and a substantial customer base,

b) 20 datasets sourced from the medical domain, characterized by a substantial number of covariates but a limited patient population.

Feature selection methods used in this study included: 3 variants of forward selection, 2 variants of a selection based on principal component analysis, a selection based on random survival forest, best subset selection and a selection based on a LASSO regularization. For details see section 5.

Additionally, the construction time needed to build a Cox model using different feature selection methods was recorded in order to assess their computational complexity. This was necessary in order to assess whether a long construction time could become a limiting factor in the model development process.

3. Literature overview: feature selection in survival modeling

It is generally recognized that there are three main categories of feature selection techniques: filter, wrapper and embedded methods (Guyon, Elisseeff 2003). Filter methods evaluate variables based on predetermined criteria, generating statistics for each feature independently. Subsequently, top-ranking features are included in the model. Thanks to the possibility of parallel calculation, these methods offer computational efficiency. Wrapper methods work on subsets of observation units derived from the original dataset of explanatory variables. A separate model is fitted to a subsample and then evaluated by a chosen statistic (describing e.g. the predictive power). Embedded methods are integrated into the process of fitting of a model itself, which is common for more complicated machine learning algorithms.

Several studies comparing different methods of variable selection for survival analysis are available in the subject literature. These analyses rely on medical data, both real and simulated. Eight out of ten articles contain numerical experiments, while the other two are meta-analyses. The summary of recent benchmark studies is presented in Table 1.

Especially in recent years several analyses have been carried out that completely omit the classical Cox model. For example, an exploration study for feature selection techniques for random survival forest was conducted by Voges, Jarren and Seifert (2023). They proposed a novel approach and applied it to several simulated datasets.

The Cox proportional hazards model stands as the conventional approach for survival analysis within medical fields. Recently, various different machine learning or artificial intelligence methods have been used more and more frequently. Despite demonstrating outcomes on par with traditional

techniques, they are often disregarded due to their opacity and limited interpretability, which are essential factors for their integration into clinical environments (Moncada-Torres et al. 2021).

Some studies compare the performance of the Cox model to more and more advanced ML-based algorithms like RSF and DeepSurv (Kar et al. 2023). The general conclusion is that more advanced machine learning methods offer slightly better performance, but the Cox model still remains a point of reference.

There are no clear guidelines for feature selection in credit risk models. In practice each financial institution implements its own methodology in line with general regulations (e.g. IFRS9, EBA guidelines or recommendations of the Polish Financial Supervision Authority). These internal methodologies are often protected by banks as a part of their competitive advantage. There are several quite recently published articles describing benchmark studies for survival analysis in the credit risk setting (e.g. Cao, Vilar, Devia 2009; Dirick, Claeskens, Baesens 2017). However, they are focused on comparing different modeling techniques rather than on feature selection methods.

Current market practice shows that banks utilize scores from PD models (classifier, a 12-month observation horizon) with appropriate transformations for the purpose of estimating a default in the lifetime horizon. Relying solely on an adjusted 12-month prediction from a PD model may lead to suboptimal results, regardless of the originally used feature selection method.

In a PD model, explanatory variables and their weights are selected to maximize performance for the annual observation window. In a lifetime horizon, however, these variables (or scores from a 12-month PD model) may perform less effectively than a dedicated Cox model (or any other survival model) with specifically chosen explanatory variables.

Typically, banks employ more or less complex filter methods for feature selection in PD models. They include, but are not limited to: forward and stepwise methods, best subset selection, elastic net regularization (including ridge and LASSO), techniques based on a single variable's predictive power and methods taking into consideration a correlation of features.

There is an abundance of censored survival data in bioinformatics. A substantial number of scientific articles deal with benchmarking feature selection methods tailored to such datasets. Since this topic has been widely investigated in the medical field and there are regulatory requirements to implement the lifetime forecast horizon in the financial sector, there arises a critical need for conducting benchmark studies focused specifically on the issue whether similar methods yield comparative results in these two industries.

4. Model construction and measurement of predictive power

One of the best and commonly used, yet simple, models appropriate for survival data is the Cox proportional hazards model (Cox 1972), given by Formula 1:

$$h(t, X) = h_0(t) \cdot \exp\left(X^T \beta\right) \tag{1}$$

$h(t, X)$ represents the hazard at time t for an individual characterized by a covariate vector X that encapsulates explanatory variables measured at the beginning of an episode. The term $h_0(t)$ signifies baseline hazard at time t, while β denotes the vector of regression parameters. Building a Cox

proportional hazards model for a given dataset involves the estimation of both the baseline hazard and the regression parameters by maximizing the partial log-likelihood function (Cox 1975).

The concordance index (C, C-index, concordance) is a widely accepted statistic measuring the model's predictive power. It was introduced by Harrell et al. (1982) and is given by Formula 2:

$$C = \frac{\sum_{i,j} \mathrm{I}_{t_j < t_i} \cdot \mathrm{I}_{sc_j < sc_i} \cdot \delta_j}{\sum_{i,j} \mathrm{I}_{t_j < t_i} \cdot \delta_j} \tag{2}$$

sc_i is model score for unit i; t_i is the unit's time of survival and δ_i is the value of the censoring variable. I is an indicator function: if $t_j < t_i$ then $\mathrm{I}_{t_j < t_i} = 1$. Else $\mathrm{I}_{t_j < t_i} = 0$. Similarly: if $sc_j > sc_i$ then $\mathrm{I}_{sc_j < sc_i} = 1$. Else $\mathrm{I}_{sc_j < sc_i} = 0$. The value of $C = 1$ corresponds to a perfect model and $C = 0.5$ represents a random prediction.

The concordance index represents the probability that, given two randomly selected individuals from the dataset, the one who experienced an event first will have a higher predicted risk. In other words, the C statistic quantifies the model's ability to correctly rank the survival times of pairs of individuals. The concordance index can be understood as a generalization of AUC (area under ROC curve) from the classification models to survival models. In this article: C_{train} represents a model's predictive power calculated on the training dataset and C_{test} – on the test dataset.

5. Feature selection methods

For the purpose of verification of the proposed hypothesis, 5 methods for variable selection were developed and adopted. Two of them were implemented in a few different variants, leading to the total of 8 different algorithms used in this study.

5.1. Forward

The forward variable selection method is a procedure used in regression analysis and statistical modeling. Starting from an empty model, individual variables are sequentially added based on their influence on the selection criterion, typically the p-value or information criteria like the Akaike (Bozdogan 1987) or Bayesian (Schwarz 1978). Additional variables are then added until specific stopping criteria are met. This method has the advantage of simplicity and interpretability but, due to its greediness, it may lead to models that are not globally optimal.

In this study, different versions of the forward method were used. Therefore, three statistics were used as the 'criterion' in the optimization process:

a) the Akaike information criterion (FA), minimized in the process,

b) the Schwarz-Bayesian information criterion (FS), minimized in the process,

c) the concordance index (FC), maximized in the process.

In other words, the goal is to add a variable resulting in the lowest possible information criterion or the highest possible concordance index of the model.

Description of the forward algorithm:

1. Estimate a model with only intercept and no explanatory variables.
2. For each feature in the set of available variables:
   a) estimate all possible Cox models containing all previously selected variables and each individual feature from the set of available variables;
   b) calculate the maximum p-value of the variables in the model (max_p_val);
   c) if max_p_val < 0.05 then proceed with the model; otherwise, discard it;
   d) calculate the 'criterion' on the training set;
   e) select the variable that optimizes the value of 'criterion';
   f) for the selected variable:
      – add it to the set of variables included in the model,
      – remove it from the set of available variables.
3. Repeat until reaching any of the following stopping criteria:
   a) desired model size;
   b) a point where no new variable can be added due to the significance level.

5.2. Principal component analysis

Principal component analysis (PCA) is a statistical technique used for reducing the size of high--dimensional data while preserving most of their variability. It operates by transforming the original variables into a new set of uncorrelated principal components, which are linear combinations of the original variables. These principal components are ordered in such a way that the first one captures the most variance in the data, and each subsequent one describes less and less variability. The goal of the PCA is to identify clusters of variables allowing a more concise representation of the data with minimal loss of information. Many feature extraction methods have been proposed based on the PCA (e.g. Al Kandari, Jolliffe 2005) but for the purpose of this analysis the following implementation was used.

1. Perform principal component analysis without considering the censoring variable. Limit the number of principal components to X or the number of principal components explaining 95% of the variability.
2. Arrange principal components in a descending order based on their eigenvalues.
3. From each component, select Y variables with the highest absolute value of the coefficient.
4. Arrange the selected variables in order of the principal component eigenvalue and coefficient absolute value. The number (index) of a variable on this list is the 'criterion' that is optimized while adding a feature to the model. In other words, the goal is to add a variable with the lowest possible index on the list.
5. Based on this set of available variables, sequentially build Cox models according to the procedure given for the forward method (see section 5.1).

X and Y are hyperparameters of the algorithm. In this study X = 50. There were two values considered for Y: 1 and 2, which resulted in two versions of the PCA method.

5.3. Random survival forest

Random forest is an ensemble method based on trees, introduced by Breiman (2001). Instead of growing a single decision or regression tree, it employs bootstrap aggregation to grow multiple trees and aggregate the results. Random forest has been extended to survival data (Ishwaran et al. 2008) resulting in random survival forest (RSF). For each split in each tree, the variable maximizing the survival difference is chosen as the best feature. Finally, the cumulative hazard function is computed using the Nelson-Aalen estimator (Aalen 1978) at each terminal node in each tree. For prediction purposes, these estimates are averaged across all trees to obtain an ensemble of cumulative hazard functions.

Permutation importance (PI) (Altmann et al. 2010) in RSF evaluates the impact of individual features on model performance by randomly shuffling their values and observing the ensuing change in prediction accuracy. The relative importance of variables in the model is obtained by permuting each feature separately and re-evaluating the model. The PI is derived from the decrease in value of performance metrics such as the concordance index.

For the purpose of this study the following implementation was used:

1. Build a RSF limiting the number of trees to Z and allowing early stopping.
2. Arrange the variables in the RSF in descending order by their permutation importance. The number (index) of a variable on this list is the 'criterion' that is optimized while adding a feature to the model. In other words, the goal is to add a variable with the lowest possible index on the list.
3. Based on this set of available variables, sequentially build Cox models according to the procedure given for the forward method (see section 5.1).

Z is a hyperparameter of the algorithm. In this study Z = 100.

5.4. Best subset

Furnival and Wilson (1974) described a method for variable selection named the branch and bound method, however, it is commonly known as best subset selection. This technique allows an efficient exploration of the large space containing a large number of available models. The search structure takes on a tree-like character. The root is comprised of a single model with all possible variables. Branches include sets of features created by removing individual predictors from more general models. Nodes of this tree assume values of the objective function for a given model. This structure allows a representation of every possible model in the search space.

In this implementation, the concordance index is the 'criterion' that is optimized while selecting the best model of each size. In other words, the goal is to select a model with the highest possible concordance index.

For each model size from 1 to the desired number of explanatory variables:

1. Use the best subset algorithm to select the top A sets of available variables.
2. Estimate all possible Cox models with given sets of explanatory variables.
3. Calculate the maximum p-value of the variables in the model (max_p_val).

4. If max_p_val < 0.05 then proceed with the model. Otherwise, discard it.
5. Calculate the 'criterion' on the training set.
6. Select the model that optimizes the value of the 'criterion',

A is a hyperparameter of the algorithm. In this study A = 10.

5.5. LASSO

Regularization involves adding penalties to various parameters of a statistical model to reduce its freedom, i.e. to avoid overfitting. In linear model regularization, a multiplicative penalty factor is applied to each model coefficient. Among various types of regularization, LASSO or L1 has a property that can shrink model coefficients to zero (Tibshirani 1996). Therefore, this method can be used as a feature selection algorithm. It has been generalized for the Cox proportional hazards model (Tibshirani 1997). In principal, every regularization technique offers a trade-off between the model's predictive power and the bias of its forecast.

Similarly to previously presented methods, the concordance index is the 'criterion' that is optimized while selecting the best model of each size.

Description of the LASSO algorithm:

1. Estimate the smallest penalty coefficient value that results in parameter shrinkage to zero for all available explanatory variables (max_alpha).
2. Generate B penalty coefficients uniformly distributed on a logarithmic scale in the range from C · log10(max_alpha) to D · log10(max_alpha).
3. For each penalty coefficient, estimate a Cox model with LASSO regularization.
4. Proceed with models containing no more than the desired number of explanatory variables.
5. For each model size:
   a) calculate the criterion on the training set,
   b) select the model that optimizes the value of the 'criterion'.

B, C and D are hyperparameters of the algorithm. In this study B = 6000, C = 0.00001 and D = 1.

6. Experiments

The study was conducted on 22 survival datasets. Two of them ('css' and 'ins') had a financial origin. They were artificially created based on a generator published by Przanowski (2011). Despite their synthetic origins, these datasets authentically mirrored interdependencies often encountered within the realm of credit risk management.

Financial datasets 'css' and 'ins' contained information on the time to default of cash and installment loans, respectively. The default is defined as a customer's failure to fulfill credit obligations (failure to pay installments) and in this study the event was identified at the moment when at least 3 installments of a loan were overdue. The beginning of an episode was defined as the moment of granting a loan. The time (length of an episode) was measured in months from granting a loan to the event or censoring. In this study the censoring could occur in the following cases:

1. The customer paid off credit obligation without delays.
2. The loan was active at the end of the data period.

In each of the financial datasets there were 212 potential explanatory variables available. Their values were measured at the beginning of an episode and could describe a loan or a customer. A single loan was treated as a unit of observation. A more detailed description of the available features is presented in Table 2.

The next 20 datasets were sourced from medical research and contained information on patient cohorts. The 'pbc' dataset was sourced from the *survival* R package[2]. Datasets 'bone_marrow' (Sikora, Wróbel, Gudyś 2018) and 's1data' (Mishra 2022) were used. The remainder (and majority) of the datasets used in this study were extracted from the open-source Python package *survset* (Drysdale 2022). The specific characteristics of the datasets can be found in the documentation of packages mentioned above.

The datasets utilized in this study were required to include a sufficient number of columns, serving as potential explanatory variables. The number of features ranged from 22 to over 15 thousand. This allowed the testing of feature selection algorithms and the identification of key factors influencing survival times across diverse datasets and domains. A summary of datasets is provided in Table 3.

Each of the datasets utilized in the study underwent the same data preparation process. First, missing values of explanatory variables were imputed with respective median values. Next, each dataset was randomly divided in a 3:1 ratio, creating separate "train" and "test" subsets. The training set was utilized to estimate models, while the test set was reserved for assessing the predictive performance of the models. This procedure ensured consistency and comparability across all datasets, enabling reliable evaluation of the models' predictive power.

For the prepared datasets, Cox models were sequentially constructed with varying sizes ranging from 1 to 10. By incrementally increasing model size, the study assessed the influence of additional variables on the predictive power of the Cox models, measured by the C-index.

The number of explanatory variables in models was limited to 10 due to the following reasons:

a) construction of bigger models required calculation times that were unacceptable (too long) given the performance of the machine used for estimation;

b) many models reached peak performance with the number of variables less than 10;

c) it was not possible to add features to models before reaching the size of 10 due to variable's insignificance (p-value > 0.05).

For detailed results see section 7.

Time of estimation was measured for each feature selection method and each model size. In particular, shorter times allow the conduct of more experiments resulting in better hyperparameter tuning of the modeling pipeline. On the other hand, the construction time can also be a limiting factor. If it exceeds a certain threshold, the study may be abandoned due to constraints arising from the schedule of the research.

All models were built on the same architecture (both hardware and software): SAS 9.4, Python 3.10, *scikit-survival* 0.22.2 (Pölsterl 2020), *survset* 0.2.6 (Drysdale 2022). The analysis of construction times focused on comparing magnitudes between different methods.

[2] Technical documentation of the *survival* R package available at: https://CRAN.R-project.org/package=survival.

7. Results

Table 4 presents the variable selection method leading to the Cox model with the highest C-index on the test dataset along with the optimal model size (number of predictors). The overfit of the model was calculated as $C_{train} - C_{test}$ and expressed in percentage points.

In the case of financial data, the best subset and forward Schwarz methods proved to be the most effective. It is also noteworthy that for financial data, feature selection algorithms yielded models of larger size compared to medical data.

As presented in Table 5, overall optimal models were constructed using the best subset and random survival forest methods, which tied for the first place in the ranking (one model selected for each dataset). Each method generated an optimal model for 6 datasets (best subset: 1 financial dataset and 5 medical ones; RSF: 6 medical datasets). Interestingly, the LASSO method yielded an overall best model for only 2 datasets (both of them medical).

It is worth analyzing which method generated an optimal model in terms of predictive power measured by C_{test} for each model size. These results are presented in Table 6.

The analysis of feature selection methods across financial and medical datasets reveals notable variations in performance. In the financial domain, the forward Schwarz and best subset methods emerge as top performers, demonstrating superior predictive power compared to others, across almost all model sizes. Conversely, in medical datasets, LASSO stands out as the clear winner, followed by random survival forest and forward Schwarz. This dominance of LASSO in medical contexts underscores its efficacy in identifying significant predictors amid high-dimensional data typical in biomedical research.

Interestingly, both the forward Akaike and forward concordance methods exhibit subpar performance across both financial and medical datasets. Moreover, PCA with 2 variables outperforms PCA with a single variable in the context of medical data.

Based on the frequency of best models of each size, the random survival forest and LASSO methods offer exceptional performance for medical datasets, affirming current literature on the subject. However, these methods fail to provide satisfactory Cox models in the financial dataset context.

Therefore, the analyzed hypothesis was disproved. Feature selection methods providing good predictive power in medical setting (LASSO and random survival forest) do not perform well for Cox models based on financial data.

Table 7 presents the average overfit for each method, by model size. A gradient scale was used for color coding. Low values were marked in blue, while high values (big overfit) were marked in red.

Overfit is practically non-existent in financial datasets, irrespective of the model size or method utilized (max overfit less than 1 p.p.). Higher values of the difference between the concordance index on train and test datasets are prevalent for medical data. In general, the greater the model size, the more overfit. However, the LASSO method generates Cox models much less prone to overfit than any other method employed in this study (with max overfit c.a. 5 p.p. at 10 explanatory variables).

The maximum acceptable level of overfit should be specified by the researcher and is unique for each specific implementation. Therefore, results presented in Table 7 do not disqualify the models obtained. They suggest, however, that medical data may inherently pose greater challenges in modeling, and models with smaller size might be preferred. The higher risk of overfit in the medical field may stem from a fewer number of observation units, compared to the financial sector. Constructing

an effective model on medical data may necessitate the application of a tailored loss function within a variable selection algorithm or the implementation of the cross-validation method to address overfit concerns.

In contrast, the absence of such requirements in financial data simplifies the potential implementation of survival modeling in banking institutions, thereby streamlining model development and deployment processes.

Figure 1 depicts construction times of models according to their size and feature selection method. Financial datasets are highlighted in shades of red and medical datasets – in blue. The aim of this figure is to illustrate general trends; hence, for clarity, a detailed legend was omitted.

For the majority of methods, including all the three variants of the forward method, LASSO and random survival forest, model building times exhibit a similar pattern: a gradual increase in logarithmic scale (base 10) with respect to model size. Feature selection methods based on PCA generally allow the creation of smaller models for medical data compared to the financial setting. Additionally, their building times follow a trend similar to the one presented by the methods listed above.

Construction times using the best subset method are notably the longest. Beyond a certain threshold, the logarithm of time increases almost linearly with the number of variables in the model, resulting in unacceptably long training times.

Based on these findings it should be stated that using the best subset method may be a limiting factor in developing a survival model due to exceptionally long estimation time increasing with the number of explanatory variables. The remaining methods used in this study did not exhibit such limitations.

8. Limitations

While this study provides valuable insights into the comparative performance of models resulting from different feature selection techniques and dataset domains, several limitations should be carefully considered. Firstly, the analysis was focused solely on Cox proportional hazards models, excluding other survival analysis methods that may offer alternative advantages. While Cox models offer great interpretability and flexibility, they may lack in predictive power compared to ML or AI algorithms. It is worth noting that easy explainability of Cox models is advantageous when it comes to meeting regulatory requirements of IFRS9.

Additionally, the choice of feature selection techniques was broad but not exhaustive, potentially overlooking other promising methodologies. As the research is constantly developing, new feature selection methods may be proposed.

Another limitation is the reliance on simulated financial datasets, which may not fully capture the complexity and nuances of real-world financial data. Overcoming this issue for a research paper would be almost impossible due to strict banking regulations related to data governance. Moreover, the medical datasets used in the study (while numerous) may not be representative of all medical domains, potentially limiting the general applicability of the findings.

Furthermore, the evaluation metric employed in the study (concordance index) provides valuable insights about predictive power but does not capture all aspects of model performance comprehensively, specifically in regards to the quality of calibration.

Individual models were not compared in terms of calibration, which refers to the alignment of the estimated probability levels with the actual event frequencies at various time points. Typically, the integrated Brier score is utilized for comparing models in this regard (Gerds, Schumacher 2006). The decision not to test the quality of calibration is motivated by the fact that the indications of a survival model can be manually calibrated to the appropriate probability values of an event at a given moment using any monotonic transformation $R \rightarrow [0,1]$. Logistic regression (Cox 1958) or isotonic regression (de Leeuw, Hornik, Mair 2009) are often utilized by banks to calibrate the score.

In the banking industry, during the model development process, generating a score with the highest possible predictive power and calibrating it to desired levels are often separated. This principle is particularly strong in business applications, where determining the appropriate ordering of observation units is more important than presenting the forecast in the form of a probability of an event. Moreover, the IFRS9 contains complex guidelines for the models' calibration, specifically separating development of the score from the development of the calibration function. Therefore, it was decided not to test the quality of the models' calibration, which was not pertinent for examination of the initial hypothesis.

The primary objective of this article is to examine the impact of variable selection methods on model predictive power. Therefore, the assumptions of the Cox model: the linear relationship between the logarithm of the hazard and the explanatory variables, and the assumption of proportional hazards were not tested. Additionally, the presence of multicollinearity among the selected variables was not assessed. Statistical significance of the parameters (at 0.05) was verified for all variable selection methods except for LASSO. This approach allows focusing specifically on the performance of various variable selection techniques and their influence on predictive power, without delving into the broader assumptions of the Cox model. It has been shown that complying with the Cox model's assumptions can be very hard or even impossible (Rizopoulos, Molenberghs, Lesaffre 2017).

A Cox model with broken assumptions offers acceptable quality of forecasts. Such a model can be treated as a good start for further analyses, e.g. manual variable selection, an ordered or multinomial version of the model or fine-tuning in general. Several approaches have been proposed: Allison (2010); Hosmer, Lemeshow, May (2008); Borucka (2017).

In this study datasets were randomly divided into train and test subsamples. Employment of cross-validation would enhance the stability and robustness of acquired results, potentially reducing the overfit for medical data.

Addressing these limitations in future studies will contribute to a more comprehensive understanding of feature selection methodologies in survival analysis and their applicability across different domains.

9. Conclusions

This study provides a comparative analysis of numerous methods of feature selection for Cox proportional hazards models across datasets from the medical and financial sectors. The main purpose of this article was to investigate whether methodologies yielding optimal models within the medical domain would exhibit superior performance across financial datasets. In particular, it was tested whether the LASSO regularization or the selection based on the random survival forest method

(generating good models for medical data, as reported in renowned literature) would offer the same performance for financial datasets. This hypothesis was hereby disproved.

Given the IFRS9 regulations applied to credit risk modeling in banks, this study shows that the best subset selection or a forward method based on the Schwarz information criterion generate Cox models with the highest predictive power for the purposes of credit risk management. Given the good performance exhibited by the best subset method on both financial and medical data, this method is recommended for the construction of survival models for purpose of credit risk management.

It has been shown that Cox models built on financial data do not exhibit overfit regardless of the method of feature selection. This presents a great chance for banks to develop survival modeling methodologies without it being critically important to employ countermeasures against overfitting. The model construction time with the best subset selection being used might be a limiting factor for a bigger number of explanatory variables.

These insights underscore the importance of tailoring feature selection strategies to the unique characteristics of each dataset and emphasize the need for further exploration to enhance the applicability and robustness of survival analysis techniques.

By incorporating survival analysis into credit risk assessment methodologies, banks can better anticipate and mitigate potential losses associated with loan portfolios. Moreover, survival modeling enables the identification of key risk factors and the development of more accurate predictive models for assessing creditworthiness. This, in turn, can lead to more informed lending decisions and improved portfolio management strategies as well as compliance with regulations. As financial institutions continue to grapple with evolving regulatory landscapes and increasing market uncertainties, the integration of survival modeling offers a valuable opportunity to bolster risk management frameworks and enhance overall operational efficiency.

References

Aalen O. (1978), Nonparametric inference for a family of counting processes, *The Annals of Statistic*s, 6(4), 701–726.

Al Kandari N.M., Jolliffe I.T. (2005), Variable selection and interpretation of correlation principal component, *Environmetrics*, 16, 659–672.

Allison P.D. (2010), *Survival Analysis Using SAS. A Practical Guide*, SAS Institute.

Altmann A., Toloşi L., Sander O., Lengauer T. (2010), Permutation importance: a corrected feature importance measure, *Bioinformatics*, 26(10), 1340–1347.

Bommert A., Welchowski T., Schmid M., Rahnenführer J. (2022), Benchmark of filter methods for feature selection in high-dimensional gene expression survival data, *Briefings in Bioinformatics*, 23(1), bbab354.

Borucka J. (2017), *Analiza i modelowanie ryzyka zachorowalności: parametryczne i semiparametryczne modele przeżycia*, Oficyna Wydawnicza SGH.

Bozdogan H. (1987), Model selection and Akaike's Information Criterion (AIC): the general theory and its analytical extensions, *Psychometrika*, 52, 345–370.

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

Cao R., Vilar J.M., Devia A. (2009), Modelling consumer credit risk via survival analysis, *SORT*, 33(1), 3–30.

Cox D.R. (1958), The regression analysis of binary sequences, *Journal of the Royal Statistical Society, Series B* (*Methodological*), 20(2), 215–242.

Cox D.R. (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*, 62(2), 269–276.

Dirick L., Claeskens G., Baesens B. (2017), Time to default in credit scoring using survival analysis: a benchmark study, *Journal of the Operational Research Society*, 68, 652–665.

Drysdale E. (2022), SurvSet: an open-source time-to-event dataset repository, *ArXiv*, abs/2203.03094.

Furnival G., Wilson R. (1974), Regressions by leaps and bounds, *Technometrics*, 16(4), 499–511.

Gerds T.A., Schumacher M. (2006), Consistent estimation of the expected brier score in general survival models with right censored event times, *Biometrical Journal*, 48(6), 1029–1040.

Guyon I., Elisseeff A. (2003), An introduction to variable and feature selection, *Journal of Machine Learning Research*, 3, 1157–1182.

Harrell Jr. F.E., Califf R.M., Pryor D.B., Lee K.L., Rosati R.A. (1982), Evaluating the yield of medical tests, *Journal of the American Medical Association*, 247(18), 2543–2546.

Heinze G., Wallisch C., Dunkler D. (2018), Variable selection – a review and recommendations for the practicing statistician, *Biometrical Journal*, 60(3), 431–449.

Herrmann M., Probst P., Hornung R., Jurinovic V., Boulesteix A.L. (2021), Large-scale benchmark study of survival prediction methods using multi-omics data, *Briefings in Bioinformatics*, 22(3), bbaa167.

Hosmer D.W., Lemeshow S., May S. (2008), *Applied Survival Analysis. Regression Modeling of Time-to--event Data*, Wiley.

Ishwaran H., Kogalur U.B., Blackstone E.H., Lauer M.S. (2008), Random survival forests, *The Annals of Applied Statistics*, 2(3), 841–860.

Kantidakis G., Putter H., Lancia C., Boer J., Braat A.E., Fiocco M. (2020), Survival prediction models since liver transplantation – comparisons between Cox models and machine learning techniques, *BMC Medical Research Methodology*, 20(1), 277.

Kar İ., Kocaman G., İbrahimov F., Enön S., Coşgun E., Elhan A.H. (2023), Comparison of deep learning--based recurrence-free survival with random survival forest and Cox proportional hazard models in Stage-I NSCLC patients, *Cancer Medicine*, 12(18), 19272–19278.

Lang M., Kotthaus H., Marwedel P., Weihs C., Rahnenführer J., Bischl B. (2015), Automatic model selection for high-dimensional survival analysis, *Journal of Statistical Computation and Simulation*, 85(1), 62–76.

de Leeuw J., Hornik K., Mair P. (2009), Isotone optimization in R: Pool-Adjacent-Violators Algorithm (PAVA) and active set methods, *Journal of Statistical Software*, 32(55), 1–24.

McWilliam A., Khalifa J., Vasquez Osorio E., Banfill K., Abravan A., Faivre-Finn C., van Herk M. (2020), Novel methodology to investigate the effect of radiation dose to heart substructures on overall survival, *International Journal of Radiation Oncology, Biology, Physics*, 108(4), 1073–1081.

Mishra S. (2022), A comparative study for time-to-event analysis and survival prediction for heart failure condition using machine learning techniques, *Journal of Electronics, Electromedical Engineering, and Medical Informatics*, 4(3), 115–134.

Moncada-Torres A., van Maaren M.C., Hendriks M.P., Siesling S., Geleijnse G. (2021), Explainable machine learning can outperform Cox regression predictions and provide insights in breast cancer survival, *Scientific Reports*, 11, 1–13.

Petersson S., Sehlstedt K. (2018), *Variable Selection Techniques for the Cox Proportional Hazards Model: A Comparative Study*, University of Gothenburg, School of Business, Economics and Law.

Pölsterl S. (2020), Scikit-survival: a library for time-to-event analysis built on top of scikit-learn, *Journal of Machine Learning Research*, 21(212), 1–6.

Przanowski K. (2011), Banking retail consumer finance data generator – credit scoring data repository, *Finansowy Kwartalnik Internetowy e-Finanse*, 9, 44–59.

Rizopoulos D., Molenberghs G., Lesaffre E.M. (2017), Dynamic predictions with time-dependent covariates in survival analysis using joint modeling and landmarking, *Biometrical Journal*, 59(9), 1261–1276.

Sauerbrei W., Perperoglou A., Schmid M., Abrahamowicz M., Becher H., Binder H., Dunkler D., Harrell Jr F.E., Royston P., Heinze G. (2020), State of the art in selection of variables and functional forms in multivariable analysis – outstanding issues, *Diagnostic and Prognostic Research*, 4, 3.

Schwarz G. (1978), Estimating the dimension of a model, *The Annals of Statistics*, 6(2), 461–464.

Sikora M., Wróbel L., Gudyś A. (2018), GuideR: a guided separate-and-conquer rule learning in classification, regression, and survival settings, *Knowledge-Based Systems*, 173, 1–14.

Spooner A., Chen E., Sowmya A., Sachdev P., Kochan N.A., Trollor J., Brodaty H. (2020), A comparison of machine learning methods for survival analysis of high-dimensional clinical data for dementia prediction, *Scientific Reports*, 10, 20410.

Tibshirani R. (1996), Regression shrinkage and selection via the Lasso, *Journal of the Royal Statistical Society*. Series B (Methodological), 58(1), 267–288.

Tibshirani R. (1997), The Lasso method for variable selection in the Cox model, *Statistics in Medicine*, 16(4), 385–395.

Vinga S. (2021), Structured sparsity regularization for analyzing high-dimensional omics data, *Briefings in Bioinformatics*, 22(1), 77–87.

Voges L.F., Jarren L.C., Seifert S. (2023), Exploitation of surrogate variables in random forests for unbiased analysis of mutual impact and importance of features, *Bioinformatics*, 39(8), btad471.

Disclaimer

The research was not directly financed. There is no conflict of interest.

Appendix

Table 1
Summary of benchmark studies

Authors (year of publication)	No. of methods	No. of datasets	Conclusion
Lang et al. (2015)	4	4	The LASSO regularization improves the Cox model performance on all datasets but to a different degree
Heinze, Wallisch, Dunkler (2017)	8	1	The backward elimination method with AIC criterion performed best for the Cox model
Petersson, Sehlstedt (2018)	15	1	All subset selection, best subset selection, backward elimination perform well for feature selection in a simulated dataset. The LASSO method performed just as well but showed shorter calculation times
Kantidakis et al. (2020)	6	1	Neural networks show better performance than random survival forests and Cox models. The best variables selection method for the Cox model was backward elimination based on the C-index
McWilliam et al. (2020)	2	1	The elastic net method with different parameterization (including ridge and LASSO methods) resulted in better feature selection than the random survival forest
Sauerbrei et al. (2020)	meta analysis		There are some papers providing general recommendations for variables selection techniques. However, the authors stress that "clear guidance is almost always impossible" and each research should be investigated separately
Spooner et al. (2020)	8	2	Cox model with variable selection based on a random survival forest performed better than the Cox model using shrinkage methods (ridge, LASSO and elastic net)
Herrmann et al. (2021)	11	18	Cox model with manually selected variables was outperformed only by a block forest method in random forest
Vinga (2021)	meta analysis		Structured regularization methods may provide better results for survival analyses suggesting that a simpler version (e.g. LASSO) may perform well
Bommert et al. (2022)	14	11	A filter method based on the feature's variance outperforms methods based on boosting algorithms and random survival forest

Table 2
Groups of potential explanatory variables in financial datasets

Type of features	Prefix in the dataset	Explanation	No. of variables	Examples
Application	APP_	Characteristics of the loan or socio--demographic features of the customer at the moment of application	12	Loan amount, customer's income
Current	ACT_	Characteristics describing the customer's current involvement with the bank	44	Number of active loans, total amount of payable installments
Behavioral 1	AGS_	Characteristics based on a history of the customer's behavior. Values are non-missing if the length of customer's relation with the bank is at least as long as the variable's summary period	78	Number of installments overdue in the last 12 months. Calculated only if the customer's relation with the bank is no shorter than 12 months
Behavioral 2	AGR_	Characteristics based on a history of the customer's behavior. Values are always non-missing, even if the length of customer's relation with the bank is shorter than the variable's summary period	78	Maximum number of days past due for all loans of the customer in the last 3 months. Calculated even if the customer's relation with the bank is shorter than 3 months

Table 3
Summary of datasets used in this study

Industry	Dataset	No. of observations	No. of variables	Censoring rate (in %)
Financial	css	46,290	212	23
Financial	ins	72,791	212	43
Medical	aids	2,139	22	24
Medical	AML_Bull	116	6,277	42
Medical	bone_marrow	187	34	55
Medical	chop	414	3,830	60
Medical	DBCD	295	4,915	73
Medical	DLBCL	240	7,392	43
Medical	gse1992	124	15,515	72
Medical	gse4335	115	12,783	67
Medical	hepatoCellular	227	30	57
Medical	mcl	92	574	30
Medical	nki70	144	71	67
Medical	nsbcd	115	549	67
Medical	dataOvarian1	912	158	40
Medical	pbc	418	20	61
Medical	phpl04K8a	442	20	47
Medical	s1data	299	11	68
Medical	smarto	3,873	15	88
Medical	supp	9,105	21	32
Medical	vdv	78	4,701	56
Medical	wpbc	198	32	76

Table 4
Feature selection methods leading to optimal Cox model on each dataset

Industry	Dataset	Method	Model size	C-index train (in %)	C-index test (in %)	Overfit (in p.p.)
Financial	css	best subset	10	58.4	58.5	0.17
Financial	ins	forward Schwarz	10	57.9	58.2	0.25
Medical	aids	best subset	7	74.4	69.7	-4.70
Medical	AML_Bull	forward Schwarz	7	83.6	78.0	-5.53
Medical	bone_marrow	PCA 2 variables	5	74.4	74.4	-0.01
Medical	chop	best subset	2	68.6	63.5	-5.17
Medical	DBCD	LASSO	10	74.6	73.8	-0.74
Medical	DLBCL	random survival forest	8	68.6	67.1	-1.50
Medical	gse1992	best subset	3	78.3	85.0	6.71
Medical	gse4335	best subset	9	87.1	82.7	-4.35
Medical	hepatoCellular	random survival forest	3	66.1	76.9	10.82
Medical	mcl	random survival forest	10	82.6	83.6	1.04
Medical	nki70	PCA 2 variables	2	69.3	77.8	8.56
Medical	nsbcd	LASSO	10	81.4	81.8	0.39
Medical	dataOvarian1	PCA 2 variables	8	62.3	64.3	1.98
Medical	pbc	forward Schwarz	5	84.7	81.6	-3.15
Medical	phpl04K8a	forward C-Index	5	64.3	70.8	6.50
Medical	s1data	random survival forest	4	73.0	72.8	-0.23
Medical	smarto	random survival forest	5	66.2	69.4	3.23
Medical	supp	best subset	5	72.8	72.1	-0.65
Medical	vdv	random survival forest	4	80.6	71.3	-9.24
Medical	wpbc	PCA 2 variables	2	68.4	63.5	-4.91

Table 5
Feature selection methods leading to optimal models – summary

Method	Number times best
Best subset	6
Random survival forest	6
PCA 2 variables	4
Forward Schwarz	3
LASSO	2
Forward C-Index	1
Forward Akaike	0
PCA 1 variable	0

Table 6
Feature selection methods leading to optimal models of each size

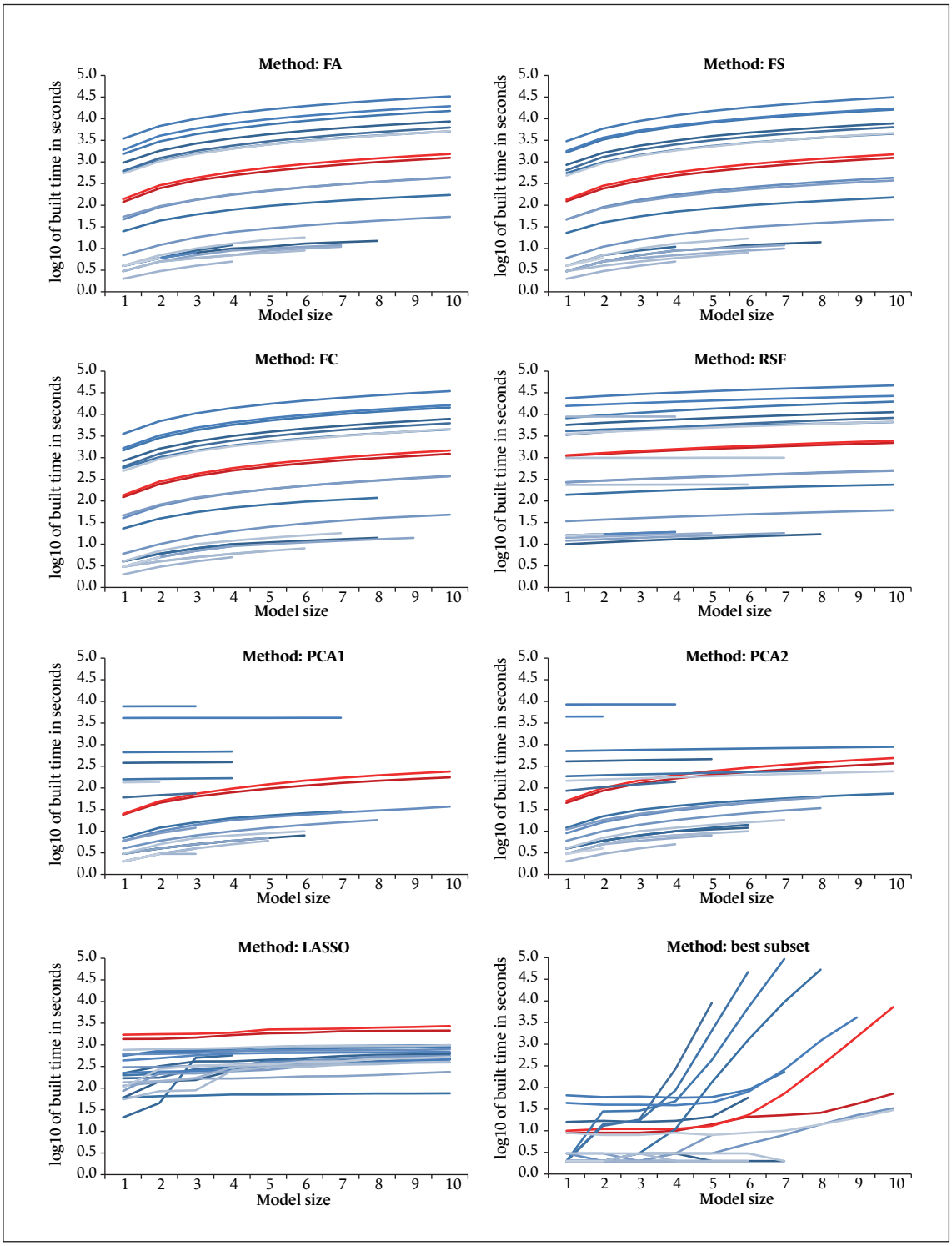
Industry	Model size	Method							
		best subset	forward Akaike	forward C-Index	forward Schwarz	LASSO	PCA 1 variable	PCA 2 variables	random survival forest
Financial	**Total**	**7**		**3**	**9**				**1**
	1				1				1
	2				2				
	3			2					
	4	1		1					
	5	1			1				
	6	1			1				
	7	1			1				
	8	1			1				
	9	1			1				
	10	1			1				
Medical	**Total**	**21**	**1**	**6**	**28**	**72**	**3**	**22**	**37**
	1	1			4	5		4	6
	2	3		1	3	4		4	5
	3	1			8	3	1	2	5
	4	2		2	3	3	1	4	5
	5	7		1	2	5	1	1	2
	6	3	1		2	9		2	2
	7	2			2	9		1	4
	8	1		1	2	10		1	3
	9	1		1	1	11		1	3
	10				1	13		2	2
	Total	**28**	**1**	**9**	**37**	**72**	**3**	**22**	**38**

Table 7
Average overfit of Cox models for all methods across all model sizes

Industry	Model size	Method							
		best subset	forward Akaike	forward C-Index	forward Schwarz	LASSO	PCA 1 variable	PCA 2 variables	random survival forest
Financial	**Total**	**0.03**	**-0.04**	**0.25**	**-0.04**	**0.31**	**0.42**	**0.25**	**0.12**
	1	0.19	0.19	0.19	0.19	0.41	0.12	0.12	0.07
	2	-0.02	-0.02	-0.02	-0.02	0.41	0.12	0.19	-0.08
	3	0.26	0.12	0.29	0.12	0.39	0.07	0.08	0.25
	4	0.23	0.05	0.30	0.05	0.41	0.53	0.09	0.19
	5	0.21	-0.06	0.27	-0.06	0.35	0.54	0.12	0.15
	6	0.07	-0.10	0.28	-0.10	0.22	0.67	0.12	0.18
	7	-0.09	-0.08	0.28	-0.08	0.23	0.59	0.45	0.12
	8	-0.21	-0.15	0.27	-0.15	0.23	0.59	0.45	0.10
	9	-0.17	-0.19	0.29	-0.19	0.20	0.51	0.40	0.13
	10	-0.19	-0.20	0.29	-0.20	0.22	0.49	0.50	0.11
Medical	**Total**	**7.88**	**12.45**	**15.60**	**12.44**	**2.95**	**2.38**	**4.46**	**9.02**
	1	2.39	3.13	7.24	3.13	1.86	-0.04	-0.84	5.71
	2	4.97	5.46	9.32	5.46	1.40	1.40	1.72	5.86
	3	6.50	6.47	10.32	6.47	1.09	2.16	3.52	7.02
	4	7.29	9.94	11.57	9.94	2.73	3.06	5.37	5.83
	5	8.02	13.61	15.08	13.61	2.59	2.99	4.74	8.56
	6	10.39	14.46	16.93	14.45	3.11	3.32	5.62	9.60
	7	13.68	17.21	18.82	17.20	3.45	5.17	8.88	10.98
	8	16.23	20.74	24.40	20.74	3.90	12.71	15.12	13.18
	9	22.63	24.22	28.35	24.22	4.56	11.75	12.31	15.70
	10	31.10	25.00	31.35	25.00	5.12	14.44	11.87	15.86

Note: low values are marked in blue; high values – in red.

Figure 1
Construction times across feature selection methods and model sizes



Note: financial datasets are in red; medical datasets – in blue.

Metody doboru zmiennych objaśniających do modelu proporcjonalnych hazardów Coxa. Analiza porównawcza dla danych z sektora finansowego i medycznego

Streszczenie

Celem niniejszego badania jest analiza porównawcza technik doboru zmiennych do modelu proporcjonalnych hazardów Coxa. Modelowanie czasu trwania jest dziedziną szczególnie rozwiniętą w naukach medycznych, w których wykorzystuje się je m.in. do badania przeżywalności pacjentów od momentu zdiagnozowania choroby lub do analizy skuteczności terapii określonymi lekami.

Wykorzystanie analizy przeżycia w sektorze finansowym jest jednak stosunkowo rzadkie. Międzynarodowy Standard Sprawozdawczości Finansowej 9 (MSSF9) narzucił na banki obowiązek tworzenia odpisów z tytułu ryzyka kredytowego. Zgodnie z MSSF9 każda ekspozycja kredytowa powinna być przyporządkowana do jednej z czterech grup (koszyków, ang. *stage*). Szczególnym przypadkiem jest koszyk drugi, gdzie oczekiwane straty kredytowe (ECL) muszą być wyznaczane w horyzoncie zapadalności kredytu. Banki samodzielnie tworzą metodyki wyliczania ECL, dopasowane do specyfiki ich działalności i jednocześnie zgodne z wytycznymi zawartymi w MSSF9.

Praktyka rynkowa pokazuje, że wymóg modelowania strat kredytowych w całym horyzoncie życia kredytu jest spełniony dzięki odpowiednim modyfikacjom punktowego wskazania modeli klasyfikacyjnych zbudowanych na 12-miesięcznym oknie obserwacji i wykorzystywanych w innych procesach decyzyjnych banku. Praktyka taka, choć zgodna z regulacjami, może prowadzić do tworzenia modeli przeżycia nieoptymalnych z punktu widzenia mocy predykcyjnej. Odrębny dobór zmiennych objaśniających i wykorzystanie modelu proporcjonalnych hazardów Coxa (jako narzędzia łatwo interpretowalnego i ugruntowanego w literaturze) umożliwia poprawę jakości modeli w banku.

Dostępne są nieliczne pozycje literatury prezentujące analizy porównawcze metod doboru zmiennych objaśniających do modeli przeżycia w obszarze medycznym. Brak jest natomiast literatury porównującej takie algorytmy jednocześnie na danych finansowych i medycznych.

W niniejszym badaniu opracowano implementacje ośmiu algorytmów doboru zmiennych do modeli Coxa. Trzy z nich bazowały na metodzie krokowej (ang. *forward*), w której zmienne dodawano sekwencyjnie, maksymalizując wybraną statystykę: kryterium informacyjne Akaike, kryterium informacyjne Schwarza lub indeks zgodności (ang. *concordance index*). Oprócz tego przeprowadzono dobór zmiennych (w dwóch wariantach), bazując na analizie głównych składowych. Dodatkowo wykorzystano metodę podziału i ograniczeń (ang. *best subset selection*), regularyzację LASSO oraz losowy las przeżycia (ang. *random survival forest*).

Badanie przeprowadzono na 22 zbiorach danych, z czego dwa pochodziły z sektora finansowego, a 20 z sektora medycznego. Dane finansowe zostały wygenerowane za pomocą symulacji i zawierały dane opisujące klientów mających kredyty gotówkowe lub ratalne. W niniejszym badaniu początek epizodu zdefiniowano jako moment udzielenia kredytu. Czas mierzono w miesiącach. Zdarzeniem kończącym epizod było niewywiązanie się klienta ze zobowiązania kredytowego (ang. *default*). Cenzurowanie mogło wystąpić w dwóch przypadkach: klient spłacił kredyt zgodnie z harmonogramem

lub kredyt był aktywny, gdy zakończono zbieranie danych. W każdym ze zbiorów danych z sektora finansowego dostępne było 212 potencjalnych zmiennych objaśniających, których wartości opisywały klienta lub umowę kredytową. Wartości zmiennych objaśniających wyznaczano w momencie startu epizodu.

Każdy zbiór podzielono losowo na część treningową i testową w celu zapewnienia oszacowania mocy predykcyjnej o odpowiedniej jakości.

Dla każdej metody i każdego zbioru danych zbudowano modele Coxa o rozmiarze nie większym niż 10 zmiennych objaśniających. Następnie modele porównano pod kątem ich mocy predykcyjnej na zbiorze testowym, mierzonej za pomocą indeksu zgodności (ang. *concordance index*).

Założenia modelu Coxa mówiące o proporcjonalności hazardów oraz o formie funkcyjnej nie były weryfikowane w niniejszym badaniu. Nie stosowano walidacji krzyżowej ze względu na długi czas estymacji modeli, jednak wykorzystanie tego mechanizmu mogłoby korzystnie wpłynąć na stabilność wyników. Poziom istotności zmiennych weryfikowano dla wszystkich metod z wyjątkiem LASSO. Kalibracja modeli nie była poddawana testom.

Dla danych finansowych najlepsze modele proporcjonalnych hazardów Coxa otrzymano po zastosowaniu selekcji krokowej bazującej na kryterium Schwarza oraz metodzie podziału i ograniczeń. W przypadku danych medycznych, zgodnie z przewidywaniami wynikającymi z przeglądu literatury, największą mocą predykcyjną odznaczały się modele Coxa z doborem zmiennych bazującym na LASSO lub na losowym lesie przeżycia. Jednocześnie obalono hipotezę mówiącą o tym, że metody dobrze działające na danych medycznych (LASSO i RSF) będą generować dobre modele zbudowane na danych finansowych. Ze względu na dobre wyniki w przypadku zarówno danych medycznych, jak i finansowych metoda podziału i ograniczeń jest rekomendowana do dalszego wykorzystania w modelach przeżycia budowanych w instytucjach finansowych.

Zauważono, że modele budowane na danych medycznych wykazują tendencję do przeuczenia. Zjawisko to nie występowało w odniesieniu do danych finansowych, niezależnie od wykorzystanej metody doboru zmiennych.

Modele porównano pod kątem czasu budowy. Zauważono, że długi czas konstrukcji modelu wykorzystującego metodę podziału i ograniczeń może stanowić czynnik zmniejszający liczbę zmiennych objaśniających.

Słowa kluczowe: analiza przeżycia, dobór zmiennych objaśniających, ryzyko kredytowe, dane finansowe, dane medyczne