SURVIVAL ANALYSIS AS A TOOL FOR BETTER PROBABILITY OF DEFAULT PREDICTION

Michal Rychnovský*

Abstract

This paper focuses on using survival analysis models in the area of credit risk and on the modelling of the probability of default (i.e. a situation where the debtor is unwilling or unable to repay the loan in time) in particular. Most of the relevant scholarly literature argues that the survival models produce similar results to the commonly used logistic regression models for the development or testing of samples. However, this paper challenges the standard performance criteria measuring precision and performs a comparison using a new prediction-based method. This method gives more weight to the predictive power of the models measured on an ex-ante validation sample rather than the standard precision of the random testing sample. This new scheme shows that the predictive power of the survival model outperforms the logistic regression model in terms of Gini and lift coefficients. This finding opens up the prospect for the survival models to be further studied and considered as relevant alternatives in financial modelling.

Keywords: probability of default, survival analysis, logistic regression, predictive power
JEL Classification: G32, G21, C58

Introduction

This paper deals with the field of consumer credit underwriting and the mathematical models that are used. While in practice, the loan approval process is usually very complicated, it could be said that the main parts of the process could be the evaluation of client's ability to repay the loan and the verification of income and other information provided. The repayment ability is examined by checking the stability and sufficiency of income to cover all expenses and by evaluating the riskiness of the client. The riskiness of the client is typically established based on the estimation of the *probability of default* (PD) conditional to the client's characteristics. *Default* is usually defined as a violation of debt contract conditions, such as the lack of will or the inability to pay the loan back. In the case of default, the creditor (e.g. a bank or other financial institution) suffers a loss. The probability of default is then usually estimated using the logistic regression models. The regression model, also called the scoring model, assigns a score to each client, which is then used as a key factor for automated approval or rejection of the loan application in the process or as one of the main inputs for the subsequent manual underwriting.

This research addresses the probability of default modelling, which has been of interest to both academics and credit risk practitioners. Besides the above mentioned logistic regression, scholarly literature has frequently discussed and tested the prospects of using new methods. Particularly after Narain (1992) proposed the application of the survival

* University of Economics, Prague, Faculty of Informatics and Statistics (michal@rychnovsky.com).

Acta Oeconomica Pragensia, 2018, 26(1), 34–46, https://doi.org/10.18267/j.aop.594

analysis theory in the probability of default modelling, there were numerous further studies of the survival analysis models in credit risk modelling, such as Banasik et al. (1999), followed by Glennon and Nigro (2005), Bellotti and Crook (2009), Cao et al. (2009), Dirick et al. (2015) and concluded by Dirick et al. (2017). The studies usually compare the methods on the development sample or on random cross validation samples. From this point of view, it has been shown by Stepanova and Thomas (2002) and Tong et al. (2012), that the survival analysis models have a similar performance to the logistic regression model in terms of precision.

This paper challenges the precision-based comparison method and focuses on the predictive performance of the models instead. For that reason, I compare the precision of the models not only on a standard random testing sample but also on an ex-ante validation sample. Thus, I aim to respond to the results of Stepanova and Thomas (2002) and Tong et al. (2012) and show that this new way of comparison can bring different results and open up prospects for the survival models to be further studied. Specifically, I follow the logic of the logistic regression model and the diversification power measures described in Rychnovský (2011), and the idea of the survival analysis model together with the data for comparison from Pazdera et al. (2009).

1. Probability of Default Models

The probability of default is usually estimated by mathematical models developed using a company's historical data. The company can look at the history of an applicant with an approved loan and evaluate their repayment history over time. In this sample, a binary target variable is created to assess the default – e.g. a client is called defaulted if he or she was more than 90 days past due (DPD) on at least one of the first 12 monthly payments. Besides the target variable, the development sample contains a set of (usually hundreds or thousands) potential explanatory variables for each client, mainly based on the data from application forms, credit bureaus, behavioural and transactional data, and other available external sources. This sample is then used to develop a precise and stable model to estimate the probability of default for new clients. This section briefly describes the standard logistic regression approach and the survival analysis approach to be compared.

1.1 Logistic Regression Model

For the logistic regression model, a sample is required that has sufficient history in order to observe a given repayment period after the loan was issued. In the above-mentioned example of 90 DPD default definition on one of the first 12 payments, we would refer to loan vintages that are at least 15 months old. Then we can take the development data sample consisting of vectors $(\boldsymbol{x}_k, y_k)$, where $\boldsymbol{x}_k$ is the vector of potential explanatory variables of the *k*-th client and $y_k = 1$ in the case of default and $y_k = 0$ otherwise. Using the logistic regression model, we can estimate the probability of default $\pi(\boldsymbol{x}_k)$ as

$$\pi(\boldsymbol{x}_k) = \frac{e^{\boldsymbol{\beta}' \boldsymbol{x}_k}}{1 + e^{\boldsymbol{\beta}' \boldsymbol{x}_k}}.$$

The parameters $\boldsymbol{\beta}$ are then estimated using the maximum likelihood method, see Lehmann and Casella (1998) and Van der Vaart (2000), and tested for significance.

The standard modelling used here usually consists of several rounds of variable categorisations, correlation adjustments, and model building using standard automated selection methods (such as forward, backward or stepwise). The final model is then tested for precision, stability, and logic. For more information regarding the logistic regression model, its parameter estimation and significance testing then see Agresti (1990) and Hosmer and Lemeshow (2000).

1.2 Repayment Survival Model

The idea of using survival analysis models for the probability of default estimation was first published by Narain (1992). This is mainly due to its advantages in addressing the censored observations that make it easily fit the profitability modelling concept.

Survival analysis deals with the modelling of the time elapsed until a particular event occurs (this is called exit or end-point), which is conditional on the specific characteristics of the subject.

Assume that X is an absolutely continuous nonnegative random variable representing the time to exit of a subject. Denote F the distribution function and f the density of X. Then we define a *hazard function* (or *intensity*) of the subject as

$$\lambda(t) = \lim_{\mathrm{h}\to 0+} \tfrac{1}{\mathrm{h}} P(t \le X < t + h | X \le t).$$

By a *survivor function* $S(t)$ (also called *survival function*) we denote the probability that the subject will not exit until time t (will survive), i.e. $S(t) = 1 - F(t)$. Using this relation, we can rewrite the hazard function into the form

$$\lambda(t) = \lim_{h\to 0+} \frac{F(t+h) - F(t)}{h} \frac{1}{S(t)} = \frac{f(t)}{S(t)} = -\frac{\mathrm{d}}{\mathrm{d}t} \ln S(t).$$

The survival function can then be expressed as

$$S(t) = 1 - F(t) = \exp\left[-\int_0^t \lambda(u)\mathrm{d}u\right].$$

More information about the survival analysis concept can be found e.g. in Kalbfleisch and Prentice (1980).

For the purpose of this research, the semi-parametric Cox model from (Cox, 1972) is used, which is based on the hazard function of subject k at time t and has the form

$$\lambda_k(t) = \lambda_0(t) e^{\boldsymbol{\beta}' \boldsymbol{x}_k},$$

where $\boldsymbol{x}_k$ is the vector of characteristics of subject k and $\boldsymbol{\beta}$ is a vector of the parameters. The function $\lambda_0(t)$ is then called the *baseline hazard function*, independent of the subject's characteristics. The Cox model and modelling is described e.g. by Therneau and Grambsch (2000) and Persson (2002).

Typically, the survival analysis models work with censoring, i.e. the fact that we do not usually have complete information about our subject – whether it had exited or not – simply because we can only observe it during a fixed time interval of length T.

During this interval, there are three possibilities of a subject status to be observed: exit at time X, no exit until time T, or the subject leaving the survey at time C before the final status could have been obtained. For further discussion on censoring and parameter estimation, see Reisnerová (2004), Kalbfleisch and Prentice (1980) and Breslow (1974).

For the repayment survival model in this paper, it is assumed that subjects are our loan clients and exits are defaults (e.g. 90 DPD after some of their instalments). Then we assume that every client would default at least once in a lifetime (either before the end of the repayment schedule – this would be a real default – or after the end of the repayment schedule – this would be a virtual default) and that the baseline hazard function is the same for all clients, i.e. that the probabilities of default of any two clients are proportional for all time intervals. This is a basic assumption that in practice is usually accepted. In reality, there are often patterns present in the loan life-cycle (e.g. a higher probability of default at the first payments followed by a better repayment moral) that are common for the loan portfolio and the information about the individual applicant is usually not strong enough to aim for modelling individual shapes of the hazard functions.

The full set of observations can then be used to define one of the following outcomes for each observation:

- Default occurred at time t, for when the client was more than 90 DPD after the instalment scheduled on time t – this is an observation with exit.
- Observation censored at time t, for when the client did not default until time t (in the case of early repayment of the loan, the original term of the loan is taken as the censoring time).

In this case, the censoring can be called non-informative (i.e. there is no relation with the default event) because it is only caused by the fact that the loan was issued later, and therefore the observation window is shorter.

Under these assumptions, we can use the Cox model to estimate the hazard function, survival function, and the vector of parameters to find the probability of default of a client until time t as

$$\pi_t(\boldsymbol{x}_k) = 1 - S_k(t).$$

Finally, the model can be examined for precision, logic and stability in a very similar manner as common scoring models. One major advantage of the survival analysis models is that they can incorporate the censored observations, and thus extend the development data sample for the most recent observations with a short history. In addition, the survival function can give the probability of default for all observed times (contrary to the logistic regression approach where only one-time horizon is used for modelling).

2. Data for Modelling

This task uses the data and its initial transformations that have been used for the project by Pazdera et al. (2009) and is a real data sample provided for research purposes by one of the largest Czech banks. The data sample consists of several data files with information about the client, application date and loan maturity, as well as the date of default according to several default definitions. In total, there are 19,139 clients with a 5.5% default rate 90 days past due.

Further on, some of the earlier transformations implemented in (Pazdera et al., 2009) are used:

- Cleaning the data in the sense of handling the obvious outliers.
- Running univariate statistics on each variable in order to find out and solve possible inconsistencies.
- Re-categorising some nominal variables.
- Omitting the correlated variables.
- Calculating the variables necessary for the model (time in months, default indicators, etc.).

The full list of variables and the transformations used for modelling are listed in Table 1.

Table 1 | List of variables and transformations used for modelling

Variable	Original format	Format used
Sex	2 categories	2 categories
Age	15 categories	5 categories
Marital status	6 categories	2 categories
Education	8 categories	3 categories
Employment status (employed since)	11 categories	3 categories
Employer	9 categories	3 categories
Housing status	6 categories	2 categories
Repayment type	5 categories	2 categories
Credit card	2 categories	2 categories
Type of employment	10 categories	3 categories
Private telephone	3 categories	2 categories
Work telephone	4 categories	2 categories
Number of dependent persons	6 categories	Excluded
Monthly income	Numeric	5 categories
Other income	Numeric	Excluded
Credit limit	Numeric	4 categories
Distribution channel	3 categories	3 categories

Source: author

For the purpose of this task, the sample was adjusted as follows: Because of the practical differences in the risk management between fix-term and revolving loans,

only the fix-term products were used – fix-term loans (i.e. the loans with the predefined instalment structure, e.g. fixed monthly payments) have a different repayment behaviour and default times than revolving loans (such as credit cards or overdrafts). Furthermore, the number of cases in the sample provided for the fix-term loans fluctuates in the early samples and after January 2006. Therefore, to achieve a robust number of observations for each month, the data sample was cleared so that it only contained vintages from the period of January 2002 to December 2005. As the sample was originally provided in 2008, the latest observations from December 2005 are mature enough to allow 24-month default measuring.

Thus, this produces a new and more homogenous sample of data that has 9,835 observations with an observable default of 90 DPD or no default on one of the first 24 payments. Two sets of data samples for two different comparisons are then prepared.

2.1 Random Sample Preparation

In the first task, the data is randomly divided into a development sample and a comparison sample in order to develop both models on the development sample and compare their diversification power to the independent validation sample. Since for the repayment survival model all the observed defaults are used with the time of default (even though the default occurred after 24 months later), it is denoted as an exit in the following text – see the sample overview in Table 2. As can be seen in this table, there are additional exits in the development sample that can be used for the survival model building. In the comparison sample, the performance is only compared to the defaults; therefore, the exits are not relevant in this case.

Table 2 | Random sample overview

Sample	Clients	Defaults	Default rate	Exits
Development	7,000	319	4.6%	500
Comparison	2,835	129	4.6%	–
Total	9,835	448	4.6%	–

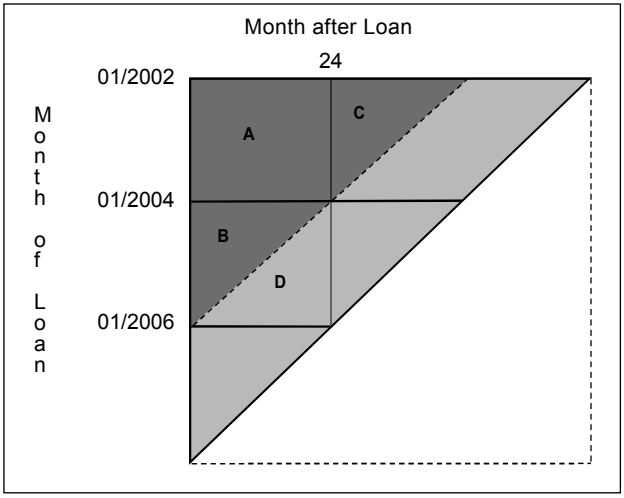
Source: author's own calculations

2.2 Progressive Time Sample Preparation

In the second task, the sample is divided by time. Assume that it is the beginning of 2006 and it is the start of the scoring model development process. When using the standard logistic regression model, only the client data from January 2002 to December 2003 can be used as a development sample (due to the fact that some time is needed to measure the defaults), whereas for the survival analysis model all the time-censored observations from 2002–2005 can be used including the partially observed vintages from 2004–2005. This is illustrated in Figure 1, where area A is the development sample for the logistic regression model, areas B and C are the additional exit observations that can be used for the repayment survival model, and area D contains information that is censored for both models and only used for the final comparison.

This is why the sample is divided into the full vintages of 2002–2003 as the development sample for the logistic regression model, the vintages from 2002–2005 time-censored to the date of 1 January 2006 as a development sample for the repayment survival model, and the full sample of 2004–2005 as the sample for comparison. For details, see Table 3.

Figure 1 | Illustration of the progressive time sample structure



Source: author's own calculations

Table 3 | Progressive time sample overview

Sample	Clients	Defaults	Default rate	Exits
Development 2002–2003	4,055	215	5.3%	279
Development 2004–2005	5,780	–	–	79
Comparison 2004–2005	5,780	233	4.0%	–
Total	9,835	448	4.6%	–

Source: author's own calculations

2.3 Comparison Measures

To compare the precision of the models, the Gini coefficient and the lift are used. The Gini coefficient is one of the most standard measures used for measuring the diversification power of a binary probability model and its precise definition together with the definitions of the distribution curve (also called the Lorenz curve or the ROC curve – from Receiver Operating Characteristic) can be found in common statistical or economic textbooks. For more information, see Hanley and McNeil (1983) and Witzany (2010).

As for the lift, the definition is taken from (Rychnovský, 2011) – the P% value of lift is defined as the ratio of the default rate for the P% worst cases divided by the default rate for the whole population. As the lift is a characteristic comparing the model performance at one relative point (e.g. a decile lift for P = 10), more information is usually taken from the complete lift curve (e.g. the values for all P ∈ [5,100]). For more information about the distribution power measures, see Řezáč et al. (2011) and Witzany (2009).

3. Results

This section presents the key results of the modelling and a comparison of the standard approach using the logistic regression model with the repayment survival model on the real Czech banking data. As mentioned earlier, the results are compared from two points of view – on the random testing sample and on the progressive time testing sample. All the calculations were made using SAS 9.4 and MS Excel.

The same 15 variables from Table 1 are used for all the models and a stepwise selection method run on significance level 0.05 for both entry and exit on the development data sample. The variables selected in the final models are summarised in Table 4.

Table 4 | List of variables included in the models

Variable	Degrees of freedom	Logistic random	Logistic prog.	Cox random	Cox prog.
Sex	1	included	included	included	included
Age	4				
Marital status	1	included	included	included	included
Education	2			included	
Employment status (employed since)	2	included	included	included	included
Employer	2	included	included	included	included
Housing status	1	included	included	included	
Repayment type	1	included		included	included
Credit card	1	included		included	included
Type employment	2				
Private telephone	1	included	included	included	included
Work telephone	1		included		
Monthly income	4				
Credit limit	3				
Distribution channel	2	included		included	included

Source: author's own calculations

The resulting model is then used on the comparison sample to calculate the score (probability of default on the first 24 payments). Finally, the results of the Gini coefficient, distribution curves and lift curves of the standard approach are compared, using the logistic regression, and the proposed repayment survival model using the Cox model, on the corresponding data samples.

3.1 Random Sample Comparison

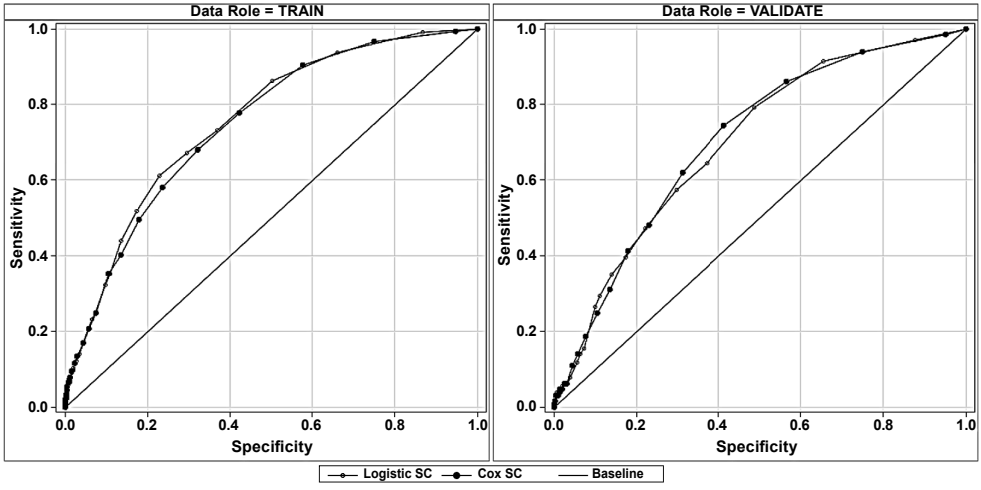
On the random sample in Table 5, it can be seen that the two models have a very similar performance in the Gini coefficient on the development (training) and comparison (validation) sample, with the Cox model being a little more stable. As for the 10% lift, it can be seen that the Cox model performs better on the training sample and worse on the testing sample. From both the distribution and lift curves in Figures 2 and 3, it can be observed that the shape of the functions is slightly different although in general it can be concluded that the performance of these models is similar.

Table 5 | Comparison of models on the random sample

Summary	Development	Comparison
Logistic regression Gini	0.51	0.39
Cox model Gini	0.50	0.40
Logistic regression lift 10%	2.73	2.78
Cox model lift 10%	2.96	2.32

Source: author's own calculations

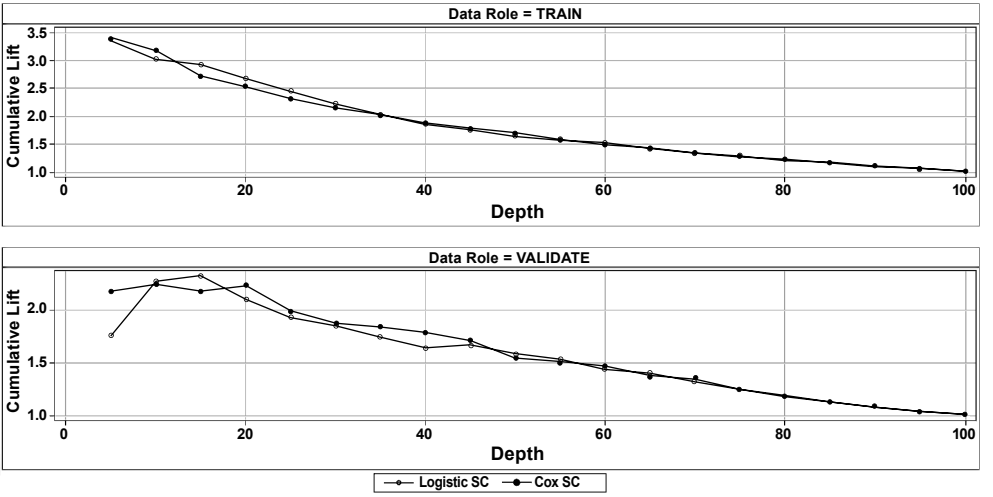
Figure 2 | Distribution curve comparison on the random sample



Source: author's own calculations

Acta Oeconomica Pragensia, 2018, 26(1), 34–46, https://doi.org/10.18267/j.aop.594

Figure 3 | Lift curve comparison on the random sample



Source: author's own calculations

3.2 Progressive Time Sample Comparison

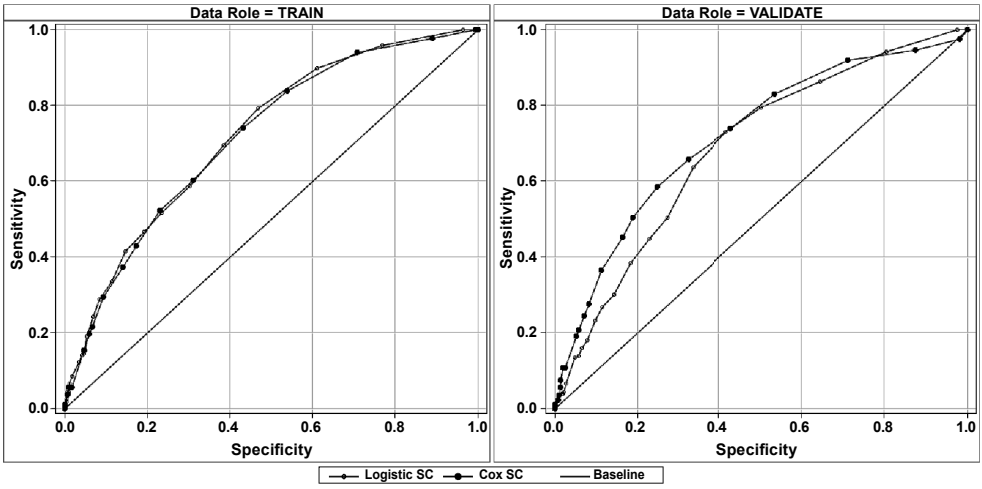
For the progressive time sample, it can be seen from Table 6 that whereas the Cox model is slightly more conservative on the 2002–2003 sample, it notably outperforms the logistic regression on the 2004–2005 sample. This is also confirmed by the distribution and lift curves in Figures 4 and 5.

Table 6 | Comparison of models on the progressive time sample

Summary	Development	Comparison
Logistic regression Gini	0.45	0.38
Cox model Gini	0.44	0.43
Logistic regression lift 10%	3.15	1.83
Cox model lift 10%	2.71	2.42

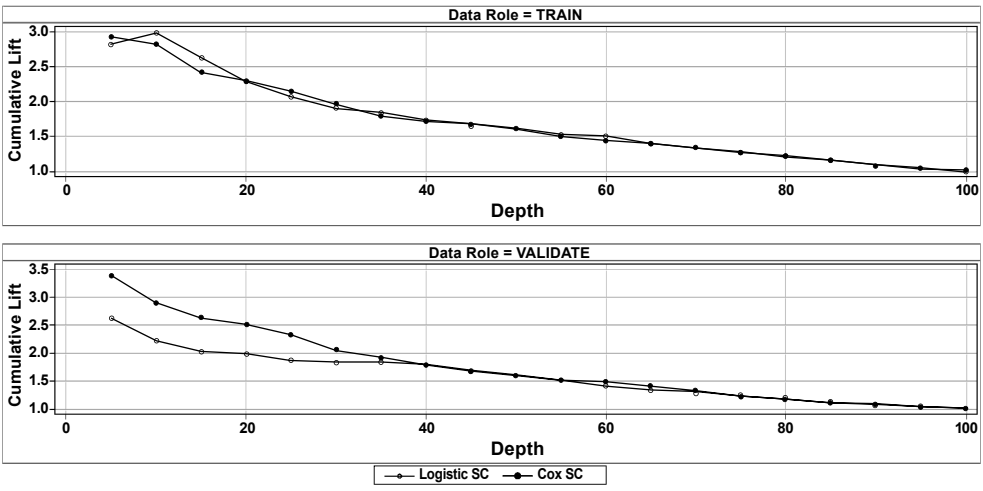
Source: author's own calculations

Figure 4 | Distribution curve comparison on the progressive time sample



Source: author's own calculations

Figure 5 | Lift curve comparison on the progressive time sample



Source: author's own calculations

Conclusions

The aim of this paper was to set new performance criteria focusing on the predictive power of models and to compare the standard logistic regression model with the alternative of the survival-based Cox model on a real sample of Czech banking data. Based on this

comparison, I concluded that in this sample both models have a similar performance on the random training and testing sample, which is in line with the existing research of Stepanova and Thomas (2002), Cao et al. (2009), Bellotti and Crook (2009) and Tong et al. (2012) and the regional specific Czech fix-term unsecured loan banking data make no exception. However, when compared with the new performance criteria measuring the predictive power of the model, the Cox model notably outperforms the logistic regression model. This is a new result contributing to the academic debate and showing the Cox model in a new light.

The survival analysis methodology was chosen on purpose, mainly because of the way it can cope with time-censored data. Therefore, it can incorporate the most recent observations into the model and thus improve its predictive power for the future. On the other hand, some of the assumptions of the model are in practice not easy to fulfil (e.g. that every client would default once – before or after the loan horizon) and the model needs to be thoroughly tested before implementation.

As can be observed from the results, in the progressive time sample the Cox model outperforms the logistic regression model in terms of the Gini coefficient and lift curves, and thus shows a better predictive power in extrapolating the last observable default vintages.

By the definition of the new performance criteria, the paper sought to contribute to the scholarly debates on mathematical modelling in finance and showed how these new criteria can change the outcomes of the probability of default models comparisons. This paves the way for further research to study the strengths and weaknesses of the survival analysis models and apply it to a broader variety of financial data to test the robustness of the claim.

As the main directions for further research, I see the testing and comparison of all the models on multiple data samples (both real and simulated) to be able to draw more general conclusions about the out-performance of the models. Moreover, the prediction horizons should be extended on the real data to understand the stability of such models in time.

References

Agresti, A. (1990). *Categorical data analysis*. New York: John Wiley & Sons, Inc.

Banasik, J., Crook, J. N. and Thomas, L. C. (1999). Not If but When Will Borrowers Default. *Journal of the Operational Research Society*, 50(12), pp. 1185-1190, https://doi.org/10.2307/3010627

Bellotti, T. and Crook, J. (2009). Credit Scoring with Macroeconomic Variables Using Survival Analysis. *Journal of the Operational Research Society*, 60(12), pp. 1699-1707, https://doi.org/10.1057/jors.2008.130

Breslow, N. (1974). Covariance Analysis of Censored Survival Data. *Biometrics*, 30(1), pp. 89-99, https://doi.org/10.2307/2529620

Cao, R., Vilar, J. M. and Devia, A. (2009). Modelling Consumer Credit Risk via Survival Analysis. *SORT*, 33(1), pp. 3-30.

Cox, D. R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2), pp. 187-220.

Dirick, L., Claeskens, G. and Baesens, B. (2015). An Akaike Information Criterion for Multiple Event Mixture Cure Models. *European Journal of Operational Research*, 241(2), pp. 449-457, https://doi.org/10.1016/j.ejor.2014.08.038

Dirick, L., Claeskens, G. and Baesens, B. (2017). Time to Default in Credit Scoring Using Survival Analysis: A Benchmark Study. *Journal of the Operational Research Society*, 68(6), pp. 652-665, https://doi.org/10.1057/s41274-016-0128-9

Glennon, D. C. and Nigro, P. (2005). Measuring the Default Risk of Small Business Loans: A Survival Analysis Approach. *Journal of Money, Credit, and Banking*, 37(5), pp. 923-947, https://doi.org/10.1353/mcb.2005.0051

Hanley, J. and McNeil, B. (1983). A Method of Comparing the Areas under Receiver Operating Characteristic Curves Derived from the Same Cases. *Radiology*, 148(3), pp. 839-843, https://doi.org/10.1148/radiology.148.3.6878708

Hosmer, D. and Lemeshow, S. (2000). *Applied logistic regression*. 2nd ed. New York: John Wiley & Sons, Inc.

Kalbfleisch, J. and Prentice, R. (1980). *The statistical analysis of failure time data.* New York: John Wiley & Sons, Inc.

Lehmann, E. and Casella, G. (1998). *Theory of point estimation*. New York: Springer Verlag.

Narain, B. (1992). Survival analysis and the credit granting decision. In: L. C. Thomas, J. Crook and D. B. Edelman, eds., *Credit Scoring and Credit Control*, 1st ed. Oxford: Oxford University Press, pp. 109-121.

Pazdera, J., Rychnovský, M. and Zahradník, P. (2009). Survival analysis in credit scoring. *Seminar on Modelling in Economics* (1 Feb. 2009). Prague: Charles University.

Persson, I. (2002). Essays on the Assumption of Proportional Hazards in Cox Regression. *Acta Universitatis Upsaliensis*. *Comprehensive Summaries of Uppsala Dissertations from the Faculty of Social Sciences 110*. Uppsala.

Reisnerová, S. (2004). Analýza přežití a Coxův model pro diskrétní čas [Survival Analysis and Cox Model for Discrete Time]. Robust, 13, pp. 339-346.

Řezáč, M. et al. (2011). How to Measure the Quality of Credit Scoring Models. *Finance a* úvěr*: Czech Journal of Economics and Finance*, 61(5), pp. 486-507.

Rychnovský, M. (2011). *Scoring Models in Finance*. MA. University of Economics, Prague.

Stepanova, M. and Thomas, L. (2002). Survival Analysis Methods for Personal Loan Data. *Operations Research*, 50(2), pp. 277-289, https://doi.org/10.1287/opre.50.2.277.426

Tong, E. N. C, Mues, C. and Thomas, L. C. (2012). Mixture Cure Models in Credit Scoring: If and When Borrowers Default. *European Journal of Operational Research*, 218(1), pp. 132-139, https://doi.org/10.1016/j.ejor.2011.10.007

Therneau, T. M. and Grambsch, P. M. (2000). *Modeling survival data: extending the Cox model*. New York: Springer Verlag.

Van Der Vaart, A. (2000). *Asymptotic statistics*. Cambridge: Cambridge University Press.

Witzany, J. (2009). Definition of default and quality of scoring functions (3 Sep. 2009). Available at SSRN: http://dx.doi.org/10.2139/ssrn.1467718.

Witzany, J. (2010). *Credit risk management and modeling*. Prague: Oeconomica.