University of Economics in Prague

Faculty of Finance and Accounting

Department of Banking and Insurance

Field of study: Financial Engineering

Advanced methods of LGD estimation

Author of the Master Thesis: Yulia Egorova

Supervisor of the Master Thesis: prof. RNDr. Jiří Witzany, Ph.D.

Date of submission: 2019

The Declaration of Authorship

I hereby declare that I carried out the master thesis «Advanced methods of LGD estimation» independently, using only the resources and literature properly marked and included in the bibliography.

Prague,

signature of the author

Acknowledgements

I would like to thank my supervisor prof. RNDr. Jiří Witzany, Ph.D. for his comments, dedicated time and support during this research.

Abstract:

This Thesis has the main aim to consider the most important Basel requirements and parameters and methods of estimation of one of them – Loss Given Default.

In Internal Rating Based Approach (IRB) frameworks banks are allowed to assess credit risk using their own models. Precise evaluation of risk parameters is important for banks to calculate regulatory capital to be able to absorb potential losses. Several methods of LGD estimation were investigated in this work in order to compare their performance.

Paper briefly describes key Basel risk parameters and explains main approaches to estimation of parameters provided by Basel Accord. Investigated models are considered from the theoretical point of view and applied on the real loan data. Results show that logistic regression and Beta regression models have better fit than other valuated models. Comparison is provided using R-squared measure.

Keywords:

LGD, recovery rate, Basel II, logistic regression, survival analysis, Cox proportional hazard model

Abstrakt:

Tato práce má hlavním cílem prozkoumat nejdůležitější požadavky Basel II a metody odhadu jednoho z nich – ztráty při selhání.

V rámci přístupu založeného na interním ratingu (IRB) mohou banky měřit své úvěrové riziko pomocí svých vlastních modelů. Přesné hodnocení parametrů rizika je důležité, aby banky byly schopné správně vytvořit svůj regulační kapitál, aby mohly absorbovat potenciální ztráty. V této práci jsou prozkoumány několik metod odhadu LGD s cílem porovnat jejich výkonnost.

Tato práce stručně popisuje klíčové rizikové parametry Basel II a vysvětluje hlavní přístupy k odhadu těchto parametrů poskytnutých Basilejskou dohodou. Modely pro odhad parametru LGD jsou popsány z teoretického hlediska a jsou aplikovány na reálná data. Výsledky ukazují, že logistická regrese a regresní Beta-model mají lepší výkonnost než ostatní hodnocené modely. Srovnání je prováděné pomocí R^2.

Klíčová slova:

LGD, míra výtěžnosti, Basel II, logistická regrese, analýza přežití, Coxův model

Contents

Introduction

Capital requirements for banks are an important condition of financial stability at a higher level in the sense that they are designed to minimize the probability of bank insolvency or at least minimize its cost. Large number of bank failures in the past have proved to be very costly in terms of taxpayers' money and have had a very negative impact on the real economy, as reflected, for example, in output decreasing and a sharp rise in unemployment. The role of capital requirements is manifested in at least two aspects: capital is a buffer covering unexpected costs and, if properly designed, encourages banks to limit the risk of their activities. Capital requirements have an impact on the return on equity, as capital is the most expensive source of financing for banks, thus potentially affecting the competitive position in the financial sector.

Against this background, and in view of the growing international mobility of capital, worldwide coordination of prudential banking supervision is essential due to the fact that it ensures equal conditions of banking operation in different countries. The Basel Accord of 1988 (Basel Committee on banking supervision (1988)) marked the beginning of a convergence of quite different approaches adopted by countries. In June 2004, the Basel Committee on banking supervision (Basel Committee on banking supervision (2006b)) published a revised version of this framework, commonly referred to as Basel II.

Basel II Accord allows financial institutions to build credit risk models to calculate their regulatory capital charge and its introduction had a great impact on the banking sector. Basel II contains modeling requirements for key risk parameters: probability of default (PD), loss given default (LGD) and exposure at default (EAD). Financial institutions can choose one of the two methods provided by The Basel Committee on Banking Supervision (BCBS): the Foundation Internal Rating-Based Approach (F-IRB), where banks can develop their own empirical model to estimate the PD and the other parameters are prescribed by regulator. Advanced Internal Rating-Based (A-IRB), suppose that banks estimate both the PD and the LGD using their own internal risk models. Under A-IRB banks are allowed to build their own quantitative models to estimate (probability of default, exposure at default, loss given default band other parameters required for calculating the RWA (risk-weighted asset).

The LGD in a broad sense can be defined as the ratio of the loss that will never be repaid to the bank to the exposure at default, or equivalently as one minus the recovery rate.

There are some alternative methods of determining LGD depending on the type of credit product and the data available. Researchers and banks often run into difficulties in measurement and the modelling of the LGD despite the fact that this definition is transparent and clear. The Basel Committee on Banking Supervision and the European Banking Authority made intensive efforts to explain in the most details and understandable way the definition of default and the scope of losses that should be considered by the banks to measure the workout LGD.

Finding and building the correct Loss Given Default model that most accurately reflects reality and predicts the future is a crucial problem in context of the importance of the LGD parameter in the Basel risk weight function and the regulatory capital for credit risk. The estimation errors in LGD have a strong impact on required capital. While probability of default is sufficiently researched topic with a lot of existing academic and practical research, LGD remains a less studied, that makes this research relevant.

The problems of the investigations conclude the following:

- examine existing models for estimation of Loss Given Default and choose the most appropriate;
- apply these methods to the real data;
- choose the methods that provide better results.

So, the main aim of this research is to find more appropriate model for loss given default estimation. Models are evaluated by comparison of their explanatory power; R^2 was chosen as an indicator of model fit.

The paper has the following structure. In Chapter 1 Basel II concept and most important requirements are explained. Also problematic of recovery rate discounting are considered. In Chapter 2 the methodology of analysis is considered from the theoretical point of view, where all of the models are described in detail. In Chapter 3 I conduct an empirical analysis and display main takeaways.

Literature Review

Empirical work, based on historical data on LGD, appeared mainly in the period 1996-2001, which was mainly due to the introduction of an advanced approach to credit risk assessment by the Basel Committee. In addition, the development of empirical research

significantly was constrained by the limitations of historical data on actual loans default losses that are only observable after the default.

Fundamental works in this field can be called researches of Altman, who have devoted several papers to the development of this topic since 1996 to the most recent in 2017. The earliest studies relied exclusively on secondary market prices of bonds or loans. Altman and Kishore (1996) estimate LGDs for several hundred defaulted senior bonds[1]. In the latest research "Intertemporal Forecasts of Defaulted Bond Recoveries and Portfolio Losses" authors estimate conditional mixtures of distributions using maximum - likelihood approach and for further comparison of intertemporal forecasting performance[2].

Special attention in the academic literature is paid to the study of factors affecting the value of LGD, mainly using econometric (regression) models in the class of parametric. To identify the parameters of the classical linear regression model, the usual least squares method (OLS) and the adjusted determination coefficient are used to assess the quality of the model. Despite the widespread use of this approach for loss modelling, it is associated with a violation of a number of assumptions of the classical linear regression.

In particular, LGD is characterized by a censored distribution, number of papers noting its bimodality with a higher concentration of observations at zero and one and a higher value of LGD during periods of economic recession. As a result, the obtained OLS parameter estimates are unreliable, and the predicted LGD values for such a model may lie beyond zero and one[3].

To overcome the above problems, different approaches are used, which include the use of the Tobit model (censored regression model) or functional transformation of the dependent variable. The logistic function and the normal distribution function are often used for functional transformation, but they do not take into account the asymmetry of the LGD distribution.

Empirical work also uses the beta distribution function [Bellotti, Crook, 2012; Stein, Gupton, 2005], and the gamma distribution. However, as was noted by Yang, Tkachenko,

[1] ALTMAN, I., KISHORE, V. M. Almost everything you wanted to know about recoveries on defaulted bonds
[2] KALOTAY, E.A., ALTMAN E.I. Intertemporal Forecasts of Defaulted Bond Recoveries and Portfolio Losses.
[3] GREENE, W.H. Econometric Analysis. s.181

the inverse transformation of the dependent variable to predict the values of the original dependent variable (LGD or RR) is usually associated with large measurement errors. [4]

A number of empirical works are devoted to the comparative analysis of predictive power of different classes of LGD models [Yashkir, Yashkir, 2013] compare predictive power of different LGD models[5]. Models include classical linear regression, Tobit model and three-level Tobit model (where three States are distinguished: LGD = 0, 0 < LGD < 1, LGD = 1), linear beta regression and its modification (inflated-beta regression model) [Cribari-Neto, 2004], as well as censored linear gamma regression.[6] Empirical results suggest that classical linear regression and beta regression have the greatest predictive power. However, the authors conclude that the predictive quality of the LGD model depends on a set of input parameters (explanatory variables) rather than on the modeling technique.

Nonparametric methods are also used. Sommers and Whittaker in modeling LGD for housing mortgage loans issued by a European Bank uses a quantile regression to predict the value of the discount in the implementation of collateral.[7]

Several papers [Qi, Yang, 2009] note a higher predictive power of nonparametric models (regression trees and neural networks) in comparison with parametric models (classical linear regression, various modifications of regression with fractional dependent variable).

Parametric models often lose to non-parametric models, which do not imply a specific distribution for LGD, in quality of predictive characteristics, parametric model's advantage is their interpretability. Qi and Zhao (2011) compare parametric and nonparametric methods (for example neural networks and regression trees with fractional response regression with using the inverse Gaussian distribution function (with beta transformation)[8]. They draw a conclusion that nonparametric methods have better

[4] YANG B.H., TKACHENKO M. Modeling of EAD and LGD: Empirical approaches and technical implementation.p. 7.

[5] YASHKIR,O., and YASHKIR, Y. Loss given default modeling: Comparative analysis. Journal of Risk Model Validation, p. 12.

[6] FERRARI, S.L.P; CRIBARI-NETO, F. Beta regression for modelling rates and proportions. p. 803

[7] SOMMERS M., WHITTAKER J. Quantile regression for modelling distributions of profit and loss. European Journal of Operational Research, p.1481

[8] QI, M., X. ZHAO. Comparison of modeling methods for loss given default, Journal of Banking and Finance, p. 2847.

performance characteristics than parametric methods but only if overfitting problem is resolved. Bastos (2010) also prefers to use nonparametric regression trees[9].

Loterman et al. (2012) compares 24 different techniques, from OLS regression to methods like robust and ridge regression, beta regression and regression splines, neural networks, regression trees and support vector regressions[10]. They find that non-linear techniques, and more specifically support vector regressions and neural networks, has significantly better performance than other traditional linear techniques. Chalupka et al (2009) use different statistical methodologies (classic linear regression models, ordinal responses models and fractional responses)[11]. They conclude that better performance have more complex models and log-log models, which imply an asymmetric response of the dependent variable. They also come to the conclusion that more important regressors that impact LGD are loan size and period of origination, length of relationship of a client with a bank and relative value of collateral.

Some studies to predict LGD besides single-stage models use more complex two-stage models. These methods are more advantageous in cases where the sample contains a large number of extreme values concentrated near 0 and 1. Bellotti and Crook (2012) provide an estimation obtained from two-stage model with using decision tree algorithm. The whole sample in this approach is splitted into three groups according to the values of LGD (0, 1, or between them).[12] Then the extremes are estimated with two logistic regression sub-models and values between are fitted by an OLS regression model. Andreeva, Crook and Yao (2017) improve this two-stage model but instead of logistic regression they apply a least squares support vector classifier. They find that this two-stage approach surpass the single-stage support vector regression model in terms of out-of-sample R-square.

Tobback et al. (2014) considered two datasets of corporate loans and home equity and also made findings of that a two-stage model, in this particular case combining linear regression and support vector regression, performs better than other techniques when

[9] BASTOS, J., Forecasting bank loans loss-given-default. Journal of Banking & Finance, Volume 34, Issue 10, October 2010, p. 2510-2517

[10] LOTERMAN, G.; BROWN, I.; MARTENS, D. Benchmarking regression algorithms for loss given default modeling, International Journal of Forecasting, p. 163.

[11] CHALUPKA R., KOPECSNI J. (2009) Modeling Bank Loan LGD of Corporate and SME Segments: A Case Study. Finance a uver – Czech Journal of Economics and Finance, 59, 2009, no.

[12] CROOK, J.; BELLOTTI, T. Loss given default models incorporating macroeconomic variables for credit cards, p. 179.

forecasting out-of-time[13]. But non-parametric regression tree in their research had better performance in out-of-sample case. Miller and Tows (2017) use multi-step estimation approach with an economic separation of the LGD determined by the workout process. Nazemi et al. (2017) implement a fuzzy fusion model which uses a function to combine the results of several base models[14]. They show that the fuzzy fusion model has higher predictive power in comparison to support vector regression models.

[13] TOBBACK, E., MARTENS, D., VAN GESTEL, T., BAESENS, B, Forecasting loss given default models: Impact of account characteristics and the macroeconomic state. Journal of the Operational Research Society. p. 387

[14] NAZEMI, A., FATEMI POUR F., HEIDENREICH, K., FABOZZI, F.J.. Fuzzy Decision Fusion Approach for Loss-Given-Default Modeling. p. 787

1 Basel II requirements and credit risk management

The Basel Accords are series of regulations aimed at the banking sector (Basel I, II and III) set by the BCBS (Basel Committee on Bank Supervision), contains capital risk, operational risk and market risk requirements and recommendations. Main purpose of the Basel is to provide requirements in regards to minimum capital of financial institutions to meet their obligations and absorb unexpected losses. The first Basel Accords appears in the 1980s it is being developed and improved constantly from that time. The Basel Committee on Bank Supervision was founded in 1974 for cooperation between its member countries on a regular base on banking supervisory matters. The first aim of The BCBS is the enhancement of "financial stability by improving supervisory knowhow and the quality of banking supervision worldwide". For now, it directs its attention to ensuring the capital adequacy of banks and monitoring of the banking system.

The most important for our purposes is the second Basel Accord, called Revised Capital Framework but better known as Basel II, which is updated version of the first accord. It focuses on three main areas: minimum capital requirements, effective use of disclosure as a lever to strengthen market discipline, internal assessment process and supervisory review of an institution's capital adequacy, and encourage sound banking practices including supervisory review. Together, these areas of focus are known as the three mutually reinforcing pillars.

Pillar I presents originate capital requirements and main innovation after its revision for credit, market and operational risk. One of them addresses to the internal or external credit ratings and its proper use for the assessment of capital requirements. It is important because of sensitivity of capital requirements to the credit quality of each specific exposure, not only to credit type. In this sense, the quality of credit starts to affect capital requirements. quality of credit quality is expressed through the probability of default and the loss given default. Moreover, for evaluating capital requirements also can be relevant indicators like maturity of credit and volume of corporate sales.

Operational risk requirements became another very important change of Basel II. Now banks have to hold capital also for operational risk. Pillar II concerns the supervision of banks. Banking supervisors need to evaluate banks' internal risk assessment

methodologies, its robustness and consistency so they have higher level of authority. Finally, Pillar III introduces requirement on the information that banks have to publish regularly. This pillar is also called the market discipline pillar. It can be noted that these capital requirements related to the quality of credit established in Basel II can have a pro-cyclical effect on the economy. this can occur because credit risk ratios tend to increase as a result of higher capital requirements when the economy is in a recession. this can occur because credit risk ratios tend to increase as a result of higher capital requirements when the economy is in a recession. It can be difficult for bank to keep enough capital in period of recession so banks tend to reduce their lending operations. This reducing can deteriorate situations and shocks in the real sector of economy. Banks capital base also can be affected by negative shocks so capital requirements is extremely important for the European banking system in view of the fact that European firms is financed usually by the bank. One of the essential requirement s of Basel II is dividing of risks and quantifying of them.

Basel II improved on Basel I, first enacted in the 1980s, by offering more complex models for calculating regulatory capital. Essentially, the accord mandates that banks holding riskier assets should be required to have more capital on hand than those maintaining safer portfolios.

The final version of the Basel II Accord was published in June 2004. This document is the result of a long communication between the banking industry, national regulators and the Basel Committee on Banking Supervision. Several proposals were released by The Committee for consultation. The Committee also offered several quantitative impact studies on its proposals, for measuring the impact of the new rules. Due to this intensive cooperation accord was significantly improved and completed.

The Basel II Accord keep the most significant moments of the Basel I Accord, including requirement for banks to hold total capital equivalent to at least 8 per cent of their total risk-weighted assets, eligible capital's definition and the basic structure of the 1996 Market Risk Amendment regarding the treatment of market risk (Basel Committee on Banking Supervision (1996)).

Hence, under Basel II, as under Basel I, the eligible capital needs to be equal to or more than 8 per cent of the risk-weighted assets, i.e., it follows the rule:

$$\frac{Eligible\ Capital}{Total\ Risk\ Weighted\ Assets} \geq 8\% \quad (1)$$

The he definition of eligible capital was only slightly modified in Basel II compared to Basel I, but the calculation of the total risk-weighted assets has been changed substantially. The total risk-weighted assets are computed as the sum of the risk-weighted assets for credit risk and a capital requirement for market risk and operational risk multiple by 12. For credit risk, a weight to each exposure is applying for calculation of the risk-weighted assets. the Committee provides a function and value of this function is the weight needed (risk weight function), where the risk drivers of each exposure are the inputs of a function. It is a general difference to the previous Accord, the fact that the weight depends on the risk drivers.

Revision of the Basel I Accord was needed by the reason of the insufficient risk sensitivity in the calculation of risk-weighted assets. Since the decision to reconsider the norms of Basel I, there has been an intention to replace the general framework of the Basel Accord with different options. Thus, according to the final version of the Basel II, banks can choose between two extensive risk-weighted asset methodologies: the Standardized approach and the Internal Ratings-based (IRB) approach.

In Internal Rating Based Approach (IRB) frameworks, banks are allowed to assess their credit risk using their own models, but the bank needs to choose which of the two options it will use, foundation or advanced.

Banks can develop their own empirical model for estimation of PD for individuals or groups of clients as part of the fundamental approach (F-IRB), but these models can only be used when approved by local regulators.

Banks can develop their own empirical model for estimation of PD for individuals or groups of clients as part of the fundamental approach (F-IRB), but these models can only be used when approved by local regulators. Regulator also prescribes LGD and other parameters that used for the risk weighted assets calculations for non-retail exposures. For retail portfolio banks have to independently estimate such parameters as probability of default, loss given default and credit conversion factor. Total required capital is a fixed percentage from these parameters.

Under A-IRB banks can evaluate use their own quantitative models to estimate PD, EAD, LGD and other parameters required for calculating the RWA, but also after approval from their local regulators. And again, required capital is obtained as a fixed percentage of the estimated RWA. For retail exposures there are no differences between a foundation and advanced approach, banks also must provide their own estimates of probability of default, exposure at default and loss given default as it was in a foundation approach. Banking corporations must always use the risk-weight functions.

The first significant difference in approach is that the Standardized approach is uses external risk assessments provided by rating agencies while the IRB supposes that bank use internal credit risk systems. Second difference is that risk weights, under the Standardized approach, are set by the Committee. These risk weights take only discrete values (as in Basel I). Under the Internal Ratings-based approach, risk weights are range of values for risk weights, that was obtained from the risk weight function which the Committee defines. To implement the IRB approach, credits should be categorized into broad classes of assets that have different characteristics of underlying risk. The classes of assets are sovereign, corporate, banks, retail and equity.

The first significant difference in approach is that the Standardized approach is uses external risk assessments provided by rating agencies while the IRB supposes that bank use internal credit risk systems. Second difference is that risk weights, under the Standardized approach, are set by the Committee. These risk weights take only discrete values (as in Basel I). Under the Internal Ratings-based approach, risk weights are range of values for risk weights, that was obtained from the risk weight function which the Committee defines. To implement the IRB approach, credits should be categorized into broad classes of assets that have different characteristics of underlying risk. The classes of assets are sovereign, corporate, banks, retail and equity.

Not all firms are included in the corporate class. Some exposures to firms can be included into a retail portfolio: small and medium sized firms (SMEs) can be classified as retail if they have exposure is smaller than 1 million euros. But it should be noticed that the regulatory treatment of small and medium sized firms categorized as corporate departs from the one applied to larger firms, according to their level of sales.

For credit risk the risk-weighted assets are estimated separately for each class of assets. The final value is obtained by applying the risk weight functions provided by the Committee to the internal estimation of risk parameters.

The Accord contains two different variants of the risk weight function one for retail portfolio and another for sovereign, bank and corporate exposures.

For sovereign, bank and corporate exposures, this function is

$$K = \left\{LGD \times N\left[\left(\frac{1}{1-R}\right)^{0.5} NI(PD) + \left(\frac{1}{1-R}\right)^{0.5} NI(0.999)\right] - LGD \times PD\right\} \times \left\{\frac{1+(M-2.5)\times b(PD)}{1-1.5b(PD)}\right\} \times 1.06 \quad (2)$$

Where R is defined as follows:

$$R = 0.12\frac{1-e^{-50PD}}{1-e^{-50}} + 0.24\left[1-\frac{1-e^{-50PD}}{1-e^{-50}}\right] - 0.04\left[1-\frac{S-5}{45}\right] \quad (3)$$

Where is b calculated using formula: $b(PD) = [0.11852 - 0.05478(PD)]^2$

S represents an annual sales function of the particular company (mln of euros),

M is the maturity of the exposure (in years),

N is the standard normal cumulative distribution, NI represents the inverse of the standard normal cumulative distribution, PD is the probability of default and LGD is the loss given default. Corporate exposures are adjusted by the sales, corresponding to the third term on the R definition. The function S equals annual sales in millions of euros if annual sales are between 5 and 50 mln euros, it equals 5 if annual sales are less than 5 mln of euros and it equals 50 if annual sales are higher than or equal to 50 million euros. Capital requirements and PD, LGD, M and Rare positively related. The positive relationship of capital requirements on maturity is dependent on the loss given default and on the level of sales. In fact, a change in M of the credit has a higher impact on capital requirements for higher values of S and LGD. Finally, the factor the Basel Committee set 1.06 as an ad-hoc adjustment in year 2004.

Capital requirements for retail exposures are:

$$K = \left\{LGD \times N\left[\left(\frac{1}{1-R}\right)^{0.5} NI(PD) + \left(\frac{R}{1-R}\right)^{0.5} NI(0.999)\right] - LGD \times PD\right\} \quad (4)$$

where

$$R = 0.03\frac{1 - e^{-35PD}}{1 - e^{-35}} + 0.16\left[1 - \frac{1 - e^{-35PD}}{1 - e^{-35}}\right] \quad (5)$$

1.1 Definition, basic and advanced models of LGD

First, we need to specify Expected Loss.

The Expected Loss of a portfolio defined as the proportion of obligors that might default within 1 year, multiplied by the outstanding exposure at default, and once more multiplied by the loss given default rate (i.e. the percentage of exposure that will not be recovered by sale of collateral etc.)[15]. Banks do not know exactly how many clients will go into default in this particular year in advance, banks also have no precise information about exact amount outstanding and the actual loss rates, so banks have to estimate average or expected values.

$$\mathrm{EL} = \mathrm{PD} * \mathrm{EAD} * \mathrm{LGD} \quad (6)$$

Where the EAD is the Exposure at Default, LGD is Loss given default and PD is probability of default.

Below we formally define these variables.

Exposure at default (EAD): this is defined as the exposure subject to be lost in the period under consideration given a default has occurred. The EAD is regarded as a random or deterministic variable, where the random element is most important for credit cards and liquidity lines.

The exposure at the time of default may be known for some products like a bond or a straight loan this is a fixed amount, but in most of the cases it isn't known because the amount varies for products like loans or credit cards. For a credit card banks know exactly only maximum outstanding exposure due to a certain credit limit. This uncertainty of the exposure at a future default is the exposure risk.

Loss Given Default (LGD): is a proportion of exposure the bank might lose in case the client defaults. It is usually expressed as a proportion of EAD and depends on the factors

[15] Basel Committee on Banking Supervision, Guidance on Paragraph 468 of the Framework Document. Basel, Basel Committee on Banking Supervision, 2005.

like type of collateral, its money value as well as the type of client and quality of work-out process. In other words, the LGD is equal to 0 % if there is no loss and the LGD is equal to 100% if the full exposure amount is lost. The LGD can also be larger than 100 % when recovery is almost equal to zero and bank has a workout costs.

For the LGD estimation are critical two definitions: clear definition of default and another one is precise definition of the loss given default concept. Depending on definition loss given default we can obtain different results. For capital calculation in accordance with Basel II standards, bank have to use the regulatory definition of default. Loss given default in accordance with this definition is based on economic loss. Loss given default in this case should reflects the recession conditions and the bank must estimate it for each placement to include all relevant risks. Historical recovery rates should be used, where applicable.

1.1.1 LGD under the foundation approach

LGD plays an important role in modern practice of banking risk management, representing one of the parameters of credit risk. In the process of internal risk management of banks, LGD is one of the key risk parameters when calculating regulatory capital in accordance with the IRB approach. The main reason for this incentive is the ability of banks to use real LGDs from their history instead of fixed regulatory LGDs. The purpose of an LGD estimate is to accurately and efficiently quantify the level of risk of recovery inherited within the default exposure.

For a senior claims on corporates, sovereigns, and banks loss given default ai set as 45% for claims that not secured by recognized collateral. For all subordinated claims (facility that is expressly subordinated to another facility) LGD is assigned as 75% for corporates, sovereigns, and banks under the foundation approach.

Most banks have already built their internal models to estimate PD, so they can apply the foundation approach. However, many banks still not able to implement fully the advanced IRB approach, because for implementation of this approach banks need to model and determine LGD.

As was mentioned above, the Loss Given Default is one of the three main components in the Basel II model. It represents the percentage of the Exposure at Default which bank expect not to be recovered by the counterparty in case of default. In other words, even if the counterparty fails to repay the amount owed, the lender will expect to receive some

percentage of the current amount owed through workout or sale of collateral. This percentage is called the recovery rate (RR), and following relation holds: RR = 1 – LGD

Where RR is defined as:

$$RR = \frac{1}{EAD}\sum_{i=1}^{n}\frac{CF_{t_i}}{(1+r)^{t_i}} \tag{7}$$

Historical data on realized losses should be used for estimation of loss given default. This value his value is related to the presence of collateral and its value, when client has no collateral, LGD of the loan is determined only by the cash flows that the he pays after default.

When calculating the LGD by the bank itself, it can use both its own data and data obtained from other external sources. In order to approve the assessment model adopted by the bank, it must show that this model is built on reliable long-term data and reflects objective reality. Estimates should be made taking into account all possible factors affecting PD, LGD and EAD. These can be direct or indirect costs, expenses associated with dictation, which the bank receives late after a default occurs and others. To direct costs include lost of capital or denied profit, indirect costs are workout costs, for example.

1.1.2 LGD under the advanced approach

Let us consider a portfolio of n credits indexed by i = 1,….,n:

Each loan is characterized by an outstanding exposure of default, a loss given default which defined as the share of EAD that bank lose in case of client's default, a probability of default that shows the probability of the default of the debtor in a one-year time horizon. The last parameter is an effective maturity M, expressed in years. The loss is then equal to

$$L = \sum_{i=1}^{n} EAD_i \times LGD_i \times D_i \tag{8}$$

where D_i is a random dummy variable equal to 1 if there is default appears before the residual maturity M_i or 0 in other case. In the advanced internal rating based approach, the regulatory capital should be big enough to cover the unexpected credit loss. The unexpected loss is defined as the difference between the 99,9% Value at Risk of the portfolio loss and the expected loss. For computation of the unexpected credit loss, the

Basel Committee applies the ASRF model. Base of this model is Merton-Vasicek "model of the firm" (Merton (1974), Vasicek (2002)), but with some additional assumptions such as the consideration that portfolios are infinitely divisible or that the risk factor is normally distributed, and that we have time horizon of one year (BCBS (2004, 2005)). So with taking into account all that assumptions we can consider the regulatory capital, can be decomposed as a sum of risk contributions (RC_i) for each particular credit which depends on the characteristics of the this particular it credit. The regulatory capital is defined by the formula:

$$RC = \sum_{i=1}^{n} RC_i \tag{9}$$

The supervisory formula for the risk contribution RC_i is given by

$$RC_i \equiv RC_i(EAD_i, PD_i, LGD_i, M_i) = EAD_i \times LGD_i \times \delta(PD_i) \times \gamma(M_i) \tag{10}$$

with

$$\delta(PD_i) = \Phi\left(\frac{\Phi^{-1}(PD_i) + \sqrt{\rho(PD_i)}\Phi^{-1}(99.9\%)}{\sqrt{1-\rho(PD_i)}}\right) - PD_i \tag{11}$$

where $\Phi(\cdot)$ id a cumulative distribution function of a standard normal distribution, ρ(PD) is a parametric decreasing function for the default correlation, and γ(M) a parametric function for the maturity adjustment. Basel Committee on Bank Supervision provides the maturity adjustment and the correlation functions that depend on the type of exposure: retail exposures, corporate, sovereign or bank exposures, The Basel II formula concentrates on the LGD, because this parameter has strong linear impact on final result and potential bank solvency strongly depends on correctness of model that estimate loss given default.

1.1.3 Probability of Default

Another important parameter of Basel Accord is the probability of default. The probability of default reflects the probability that the default will occur in a certain time horizon (usually 12 months). The most common definition is delay in payments for at least 3 months. The Basel section relating to Internal Rating Based Approach says, that each PD estimate should represent a conservative view of the long-term average PD for

the class and, therefore, should be based on historical experience and empirical data. A conservative view represents a cautious or a non-optimistic view. Moreover, in Section 444 to 485 in Basel II, it is stated that PD estimates must be calculated as a long-run averages of one-year default rates for each grade, and that internal estimates of PD must incorporate all relevant information and data and should be based on historical and empirical evidence but not only on subjective or judgmental considerations.

When a default occurs a certain portion of the outstanding amount is recovered. This is called the recovery rate and it usually has to be calculated on the facility level.

1.1.4 Exposure at Default

Another key risk parameter is exposure at default. The amount which the bank is expected lose in the event of an obligor defaulting represents the EAD i.e. exposure at default is the predicted amount of loss a bank may be exposed to when a debtor defaults on a loan. Since default occurs at an unknown future date, this amount couldn't be defined in advanced and should be estimated by the bank. Mechanics of EAD calculation also depends on approach implemented by the bank: advanced or foundation. Under foundation approach EAD is calculated using the credit conversion factor (CCF) method, where the CCFs are provided by the Basel guidelines; collaterals, guaranties or security are not taken into account while estimating EAD. For advanced approach banks are allowed to use their own models, CCFs are not provided by BCBS and have to be calculated.

1.1.5 Recovery rate

Loss Given Default and the recovery rate add up to one, therefore, the growth of one indicator leads to a decrease in the second. The rate of recovery is the percentage of the customer's exposure at the time of default that the customer has repaid during the workout process. This value is calculated on the basis of historical data of the bank and then these obtained results can be extrapolated to the whole existing portfolio of bank loans.

For market instruments the recovery rate shows the degree to which the creditor recovers the principal and accrued interest due on a defaulted debt. For the remaining debts, recovery rate consists of the amount of payments received during the workout process. In addition to incoming payments, the workout approach should also take into account the costs incurred by the bank, such as administrative and legal costs, that can be quite significant. We can express recovery rate by the formula:

$$RR = \frac{1}{EAD}\sum_{i=1}^{n}\frac{CF_{t_i}}{(1+r)^{t_i}} \tag{12}$$

Discount rate can be based on a measure of the *RR* systematic risk and a general price of risk. [16]

Recovery rates and aggregate default rate usually have inverse relationship. It can be explained by the fact, that they are both are strongly influenced by the economy. For example, the same adverse economic conditions that cause defaults to rise—such as a recession—can cause recoveries to fall.

To apply the obtained results to the existing portfolio of non-default loans, for this the bank needs to estimate two parameters - the probability of default and the loss of default in the 12-month or longer horizon.

1.2 Discounting

Since the repayment of default loans can take a long time, it is necessary to discount cash flows to the total period, the most natural of which is the moment of default. The big question in this regard is the discount rate that applies to cash flows after default. This problem has not been fully resolved to date, as Bank supervisors, practitioners and scientists have not been able to come to a General conclusion about what interest rate should be used in order to obtain an estimate of the real economic losses resulting from the default. Majority of banks' loan portfolios necessitates an actuarial approach that uses a punitive (or risk-adjusted) discount rate, but for some portfolios bank can use rate which was derived from observing the market price of defaulted debt.

Recovery cash flows are uncertain in a future and contain undiversified risk away so net present value of recovery must reflect the time value of money and corresponding risk premium for this type of risk. Risk premiums for the estimation of LGDs should be consistent with economic downturn conditions, the bank should take into consideration that defaults arise during the economic downturn so the uncertainties in recovery cash flows also can arise. If bank has no uncertainty about future recovery flows (e.g., recoveries derived from cash collateral), calculations should reflect only the time value

[16] WITZANY, J. Unexpected recovery risk and LGD discount rate determination, European financial and accounting journal. 2009, vol. 4, no. 1, p. 67.

of money, and a risk-free discount rate is appropriate." Prudential regulators allow the use of different discount rates in the LGD calculation if they include the time value of money and a systematic risk's premium. Accounting standards in contrast require the effective rate (which is the contract rate usually).

$$RR = \frac{1}{EAD}\sum_{i=1}^{n}\frac{CF_{t_i}}{(1+r)^{t-T_D}} \tag{13}$$

Where $r = r_t + \delta$ and δ is a risk premium.

Bank has to select the one discount rate to make flows at different times comparable. In the literature, different approaches to discounting are proposed.

Loan contractual rate: this approach assumes that the flows recovered after default of the client should be discounted at the contract rate which was defined at the beginning of the relationship between the bank and the client or at the last contractual rate renegotiated with him. Implementation of this approach may be considered acceptable only if there is reason to believe that this rate correctly identifies the opportunity cost of the missing recovery flows; it is assumed, then, that the appearance of the insolvency event does not modify the risk of the operation. To applying of this approach, the bank must have full information about the client because any differences in the stipulated contracts have significant effect on the capacity to renegotiate the rates and, therefore, on their time for development. So we can conclude that the bank should not use mean or aggregate rates for the LGD estimation. Moreover, should be constructed internally and it should primarily contain customer information collected directly from internal sources.

The problem with this method is that it doesn't separate required returns before default and required returns after default in the case that payments received from the liquidation of assets. As discussed above, default may result in a change in the nature of the financial claim creating a new instrument where the bank is a direct investor in the recoverable assets. Post default the required rate of return will vary depending upon the systematic risk of the financial claim. An appropriate discount rate may therefore be less than the contractual rate which includes compensation for the expected reduction in cash flows relative to promised payments.

Second possibility: bank weighted average cost of capital (WACC): applying of a weighted cost of capital as a discount rate was proposed by Witzany (2009) and Jensen (2015)[17]. Weights are obtained from equity and debt funding of a bank in market value terms proportionally. Some sources claim that the bank funding costs do not reflect the risk profile of individual credit exposures with regard to resolution risk given loan default. Furthermore, it is very difficult to determine market funding costs for distressed/defaulted assets. This approach is potentially one of the preferable if not to take into consideration this difficulty. Nevertheless, bank funding costs may provide a basis for a reasonable discount rate, but under the assumption that regulatory capital is a reasonable measure for systematic risk and that post-default capital and debt ratios are consistent with the ones of other loan instruments. Structurally, this approach combines bank-level funding costs and loan-level funding ratios.

Lender's cost of equity. This approach implies that the cost of recapitalizing the bank's balance sheet is covered by the shareholder's capital. This method mistakenly replaces the systematic risk of the defaulted debt with the risk of the bank. The two are different investments and would result in the LGD rate varying with the bank's leverage and risk premium. Valuing an assets should be made from the prospective of a market clearing price considering what a buyer will pay and not solely by what the seller wishes to recoup to repair their balance sheet.

Risk-free rate: Gordy M.B., Carey M.J. (2004) in their work has a hypothesis that LGD risk can be fully diversified and banks can discount recovery cash flows at the risk-free rate[18]. One of the reasons, why the intermediary can choose the risk-free rate for discounting of repayments is the fact, that it is very difficult to identify the possible rate identical to the particular product. Thus, the main problem with this approach is how to identify the reference market and the most appropriate available rate for the risk-free activity yield (Unal, Madan and Guntay, 2003). It is impossible after a default, to accurately predict ex ante amounts and dates of future recovery payment flows, as was the case before the default, which leads to an increase the variability of the repayment flows tied to the financing that is paid out. Even if it is assumed that the risk-free rate is a suitable value for discounting future flows received by the bank before the appearance

[17] WITZANY, J. Unexpected Recovery Risk and LGD Discount Rate Determination, European Financial and Accounting Journal. p. 68.
[18] GORDY, M.B.; CAREY, M.J. Measuring systematic risk in recoveries on defaulted debt: firm-level ultimate LGD. p. 114.

of default, after default the cash flows no longer have the original characteristics and the application of this rate may be incorrect.

LGD can be underestimated when use the risk-free rate because the current value of the sum of future recovery flows do not consider level of uncertain that characterizes the recovery flows. It also can lead to underestimation of the loss in cases of insolvency because a yield of financials provided by banks is always higher than the risk-free rate, because it has a non-zero risk of loss.

A more reliable solution may be to use the discount rate adjusted for the estimated risk calculated as the risk-free risk rate increased by the risk spread (Maclachlan, 2004). This approach suggests the possibility of determining an index that represents market risk for all debtors considered in an LGD assessment (Duellmann and Trapp, 2005). Analysis using this approach is usually based on indicators related to the average behavior of the default bonds that considered to be a proxy for the market index (Altman, Brady, Resti and Sironi, 2005) or a proxy for economic growth due to the strict relationship between the LGD and the business cycle (Frye, 2000). [19] Recent studies have shown that recovery rates systematically reflect current economic conditions (Gupton and Stein (2002)) [20]. Thus, the insolvent debt is likely to have a discount rate exceeding the risk-free rate, except in cases where the bank expects to receive payments through waiting for payments from the liquidation of cash collateral.

Another widely used method is the technique of Gibilaro et al (2007) who have tried to the estimate risk premium component through the estimation of a monofactorial rate based on market index or macroeconomic variables. That rate adds to a risk-free rate. In 2011 they also tried to evolve their model that estimates a multifactorial rate obtaining LGDs with a low volatility.

Here we have briefly considered the main provisions of the Basel II and the most important its indicators. Basel Accord has been developed in 80th years and for today contains capital risk, operational risk and market risk requirements and recommendations. Main purpose of the Basel is to provide requirements in regards to minimum capital of

[19] GUPTON, G. M.; STEIN R.M. (2002). "LossCalc™ : Model for Predicting Loss Given Default (LGD)." Moody's Investors Service.

[20] MACLACHLAN I. (2004) Choosing the discount factor for estimating economic LGD in Altman E., Resti A. and Sironi A. (eds), Recovery risk. The next challenge in credit risk management.

financial institutions to meet their obligations and absorb unexpected losses. LGD is one of the most important parameters used in the calculations of the Basel risk weight function and the regulatory capital for credit risk. When estimating LGD, it is necessary to accurately quantify the level of potential recoveries obtained after the default.

Currently, Basel offers several options for modelling the main risk parameters, the preferred of which is the one, where the bank internally calculate parameters since this method is more reliable and more accurately reflects the existing risks of a particular bank. At the same time, the implementation of this method is associated with certain difficulties. Currently, there is a broad theoretical and practical base of recommendations for calculating the probability of default, while for the loss of default there is no structured guide to the selection and comparison of models.

2 Methodology

The Basel Agreement II, adopted by the European banking system on January 1, 2007, emphasized the particular role of the LGD in calculating core indicators. The concept of LGD in the Basel Agreement 2 is quite close to that used by researchers and practitioners: it can be defined as the proportion of default exposure that will never be received by the lender. The level of efficiency (in terms of costs and time) of the workout department can significantly affect the LGD and should be reflected in the estimates used to assess recovery risk on future defaulters. Thus, the improvement of the collection procedure may lead to a reduction in empirical LGD and subsequently to a reduction in capital requirements for the following years.

The most frequently and successfully applied approach among banks is the workout approach. It is believed that this approach should reflect real risks better than all the other methods and be more representative because it is based on real historical data on payments on default loans and bank's own experience so can provide better estimates. market and implied market LGD methodology can also be used, but they have many restrictions on its application, for example, they can be used only in the case of absolute market liquidity and only for retail portfolio, but only if all requirements for estimating probability of default are met by the bank.

In this work was chosen a workout approach, where loss includes all the relevant costs subsequent upon collection process, and all the cash flows are discounted. The workout LGD calculation consists in the calculation of empirical loss rates through the observation of each charge-off at the end of recovery process, according to the following formula:

$$LGD = 1 - RR = \frac{1}{EAD}\sum_{i=1}^{n}\frac{recovery\ cash\ flow\ -\ costs}{(1+r_t)^{t-T_D}} \tag{14}$$

The general purpose of the internal LGD models lays in estimating of the LGD for the credits which are currently in the bank's portfolio and which are not in a default in a current moment. These models are based on a sample of defaulted credits for which bank has information about the ex-post workout LGD. By identifying the main characteristics of these contracts and the key factors of the recovery rates, it is then possible to forecast the LGD for the non-defaulted credits.

How should LGD models be compared? Many banks today apply the benchmarking method which simply splits a sample of defaulted credits in a training set and a test set, estimates the competing models on the training set and then, evaluates the LGD forecasts on the test set with standard statistical criteria such as the mean square error (MSE) or the mean absolute error (MAE).

2.1 Goodness of Fit Measures

Standard goodness of fit measure is represented by weighted R-squared, given the formula:

$$R^2 = 1 - \frac{\sum EAD(i) \cdot \left(LGD(i) - \hat{L}(i)\right)^2}{\sum EAD(i) \cdot (LGD(i) - \mu)^2} \tag{15}$$

The indicator $R^2 = R^2(D, \mu)$ depends on the set of defaulted accounts used and on the mean μ. μ is defined as:

$$\mu = \frac{\sum EAD(a) \cdot LGD(a)}{\sum EAD(a)} \tag{16}$$

There are several other ways to compare models. The simplest and most commonly used indicator is the Mean Squared Error. It is defined by the formula:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{17}$$

where y_i is the actual expected output and $\hat{y}_i$ is the model's prediction.

MSE basically measures average squared error of predictions. For each point, it calculates square difference between the predictions and the target and then average those values. It is probably the least useful indicator, but as was mentioned above, the simplest.

Higher MSE values indicate worse models. This indicator is always non-negative, since all individual prediction errors are squared before their summation. However, for the ideal model, the value will be zero.

Another statistic that used for model estimation is Root Mean Square Error (RMSE). The RMSE is the square root of the variance of the residuals. It indicates the absolute fit of model's predictions to the observed data in each point. In contrast with R-squared, Root Mean Square Error is an absolute measure. As the square root of a variance, RMSE can

be interpreted as the standard deviation of the unexplained variance, and has the useful property of being in the same units as the response variable.

The lower RMSE value, the better the model is. Root Mean Square Error is a good measure of how accurately the model predicts the response, and it is the most important criterion for fit if the main purpose of the model is prediction.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} = \sqrt{MSE} \tag{18}$$

What is common between RMSE and MSE? First, they are similar in terms of their minimizers. Square root is a non-decreasing function, so every minimizer of MSE is also a minimizer for RMSE and vice versa. For example, if we have two sets of predictions, A and B, and we know that MSE of B is lower than MSE of A, then we can say that RMSE of A is greater RMSE of B. And it also works in the opposite direction

$$MSE(a) > MSE(b) <=> RMSE(a) > RMSE(b) \tag{19}$$

It means that, even if the target metric is RMSE, we also can compare our models using MSE, since MSE will order the models in the same way as RMSE.

The next way to compare models is Mean Absolute Error (MAE). MAE measures the average magnitude of the errors in a set of predictions, without taking into account their direction. It's the average over the test sample of the absolute are independent variables which describe differences between prediction and actual observation where all individual differences have equal weight.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{20}$$

MAE as well as RMSE can range from 0 to ∞.

The LGD models, analyzed and compared in this paper, are based on several different regression algorithms. As outlined above, I compare estimation results for different parametric and non-parametric methods. Seven modelling algorithms are considered: linear and logistic regression, Tobit model, regression tree model, Beta regression model,

Zero-One Inflated Beta model and survival analysis represented by the Cox and modified Cox regressions. A short description of the models is below.

2.2 Linear and Logistic Regression

Linear regression is the most obvious predictive model to use for recovery rate (RR) modelling, and it is also widely used in other financial area for prediction. Formally, linear regression model fits a response variable y to a function of regressor variables $x_1, x_2, \ldots x_m$ and parameters. The general linear regression model has the form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots + \beta_m x_m + \varepsilon \quad (21)$$

Where in this case y is the recovery rate or recovery amount,

βs are unknown parameters,

$x_1, x_2, \ldots, x_m$ are explanatory variables that are client and credit characteristics and ε is a random error term.

It assumes in linear regressions that the mean of the error component (random variable ε) is zero and each error component approximately has a normal distribution. However, the distribution of recovery rate tends to have a shape with prevailing zero and ones, so the error component of linear regression model for predicting recovery rate does not satisfy these assumptions[21].

The second possibility we will explore is the logistic regression based on the idea dividing the observed and future LGDs on "low" and "high" values. We now have to define the threshold. This point is usually defined as weighted average value among a pool that weighted by exposure at default. Several studies use another technique, where each observation is divided into two dummy variables, and one of them has value one, another one - zero - with appropriate weights defined by loss given default and recovery rate, i.e. observations with one have weight of loss given default, with zero weight one minus LGD. Let $l \in (0,1)$ be a threshold and define an LGD value to be "low" if LGD is lower than threshold. Hence for $a \in A$ we have the indicator function $low(a) \in \{0,1\}$ and for a non-defaulted receivables we want to find the logistic function

[21] Zhang J., Lyn C. T. (2012) Comparison of linear regression and survival analysis using single and mixture distribution approaches in modelling LGD. *International Journal of Forecasting*, 28 (1), p. 209.

$$\pi(\alpha) = \frac{exp(\boldsymbol{x}(\alpha)'\boldsymbol{\beta})}{1 + exp(\boldsymbol{x}(\alpha)'\boldsymbol{\beta})} \tag{22}$$

estimating the probability that the loss will be "low" if the account defaults. To estimate the ex ante *LGD* we combine appropriately the *EAD* weighted mean of low observed *LGD*s and high observed *LGD*s, i.e.

$$L(\alpha) = \pi(\alpha) \cdot \mu_{low} + (1 - \pi(\alpha)) \cdot \mu_{high} \tag{23}$$

where

$$\mu_{low} = \frac{\sum_{\alpha \in A, low(\alpha)} EAD(\alpha) \cdot LGD(\alpha)}{\sum_{\alpha \in A, low(\alpha)} EAD(\alpha)} \tag{24}$$

and

$$\mu_{high} = \frac{\sum_{\alpha \in A, -low(\alpha)} EAD(\alpha) \cdot LGD(\alpha)}{\sum_{\alpha \in A, -low(\alpha)} EAD(\alpha)} \tag{25}$$

The vector of parameters $\boldsymbol{\beta}$ is obtained by maximizing the likelihood

$$L = \prod_{\alpha \in A} \pi(\alpha)^{low(\alpha) \cdot EAD(\alpha)} (1 - \pi(\alpha))^{(1 - low(\alpha)) \cdot EAD(\alpha)} \tag{26}$$

2.3 Regression trees

Regression trees are a useful way of data analysis from the point of view of their effectiveness, clarity, and interpretability of results. It can indicate factors that are particularly appropriate for a model in terms of explanatory power and variance from statistical prospective. This method can be used as a first step in building of a model or as a final visualization.

Regression trees refer to nonparametric and nonlinear prediction models, initially introduced by Breiman et al. (1984)[22]. Similar to other regression methods, they can be applied to analyze the underlying dataset and to predict the (numeric) dependent variable.

In contrast with parametric methods, such as linear or logistic regressions, regression trees make no ex-ante assumptions about the distribution of the underlying data, and no functional relationship is specified.

[22] BRIEMAN, L., FRIEDMAN, J. H., OLSHEN, R. A., STONE, C. J. Classification and regression trees.

The estimation uses a greedy search algorithm by which the initial data is recursively divided into smaller disjoint subsets[23]. The regression tree model is constructed by a number of consecutive if-then logical conditions, for example, imagine that we have bank loans data with a set of explanatory variables and a dependent variable (loss given default or recoveries). In the particular tree child nodes are created from the root node by iterating through all possible binary splits of all available variables from the original set to the one node that minimize the intra-subset variation of the dependent variable. This allows reduce the heterogeneity of the target variable step-by-step from one node to the next. This procedure is repeated until further reduction of the variation of the target variable becomes unattainable. Decreasing of the variance of the target variable is measured by the "standard deviation reduction'.

$$SDR = \sigma(T) - \frac{m(T_1)}{m(T)}\sigma(T_1) - \frac{m(T_2)}{m(T)}\sigma(T_2) \tag{27}$$

where m is mean and σ is standard deviation of the target variable in the set. T is the set of observations in the parent node and T1 and T2 are the set of observations in the daughter nodes that result from splitting the parent node according to the optimal attribute. The splitting originates from the root node and goes down the endpoints through all data. The predicted values are determined by the average value of the target variable for the set of observations in each leaf. The predicted values are always between 0 and 1 because they are given by the average LGD (resp. recovery), so regression trees are particularly suitable for loss given default modeling.

There are various algorithmic approaches for variable selection; two frequently applicable greedy algorithms are forward selection and backward elimination.

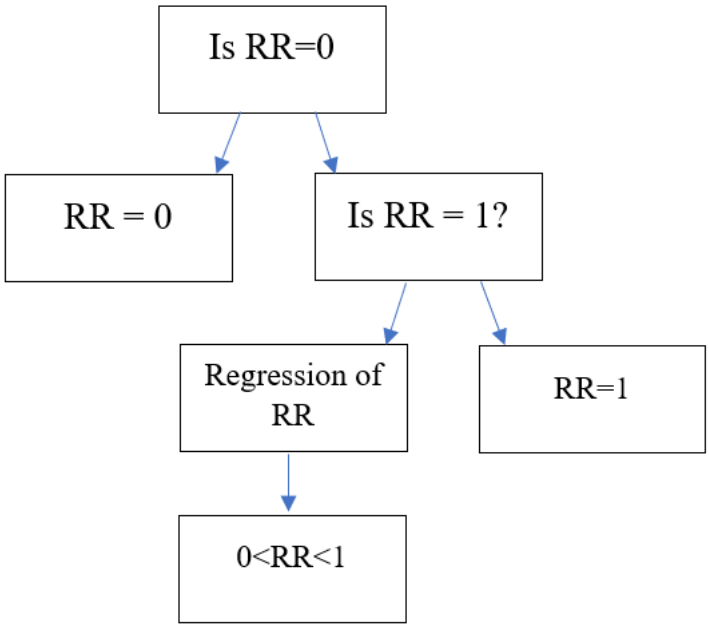
One of the problems of using this method is overfitting. Quite often final trees are massive and have good accuracy on original data but poor predictive power on new data. Efficiency improvement can be achieved by 'pruning' the tree after the main growth process. The pruning procedure is to compare the expected variance of the target variable in each node for unobserved data with the previous one. 'Pruning' occurs if the subtree variance is greater than the parent node variance and the procedure is repeated until the error is corrected.

[23] BRIEMAN, L., FRIEDMAN, J. H., OLSHEN, R. A., STONE, C. J. Classification and regression trees.

The strength of this method is that the data is collected in a node and the possible outliers have no effect on the splitting. In addition, the resulting trees are invariant under the monotone transformation of independent variables. Regression trees are particularly useful for problems with higher dimensionality because they are able to find accurate with only a few most important variables.

Regression trees are accessible for interpretation and produce results that are comparable to other more complex methods in many applications.

Decision trees are also often used in research in this field of study. The decision tree model uses two logistic regression sub-models to model the special cases of no recovery and full recovery, i.e. LGD = 0 and 1 respectively, as binary classification problems. Then, if 0<LGD<1, a regression model is used, either OLS or Least Absolute Value. This decision tree model is illustrated in Figure 1.



Source: own illustration
Figure 1. Decision tree for LGD.

This is model also good at modelling LGD due large number of boundary cases which allow to naturally approach the problem as a hybrid of classification and regression problems. This approach is meaningful since there may be special conditions which would make a customer pay back the full amount of debt or to pay back nothing, rather than a portion.

2.4 Censored Linear Regression (Tobit) Model

To take into account that LGDs are bounded at 1 in the higher domain we use the Tobit regression model.

The Tobit model, also called a censored regression model, is designed to estimate linear relationships between variables when there is either left- or right-bounds in the dependent variable (also known as censoring from below and above, respectively). Censoring from above can occur when cases with a value at or above some threshold, all take on the value of that threshold, so that the true value might be equal to the threshold, but it might also be higher. In the case of censoring from below, values those that fall at or below some threshold are censored.

When the data are censored, variation in the observed dependent variable underestimates impact of the regressors on the ”real” dependent variable, so standard ordinary least squares regression using censored data are usually biased .

If β and x_i denote the vector of coefficients, respectively, the vector of explanatory variables for observation i, ε_i the error term, and assuming that the latent variable for LGD is given by:

$$LGD_i^* = \beta' * x_i + \varepsilon_i \tag{28}$$

And $\varepsilon_i | x_i \sim N(0, \sigma^2)$. Then

$$LGD_i \sim \begin{cases} LGD_i^* & -\infty < LGD_i^* < 1 \\ 1 & LGD_i^* \geq 1 \end{cases} \tag{29}$$

The Tobit model is estimated using maximum likelihood methods.

A likelihood function can be constructed, assuming the distribution of the residuals conditional on $\boldsymbol{x}$ is normal. Maximum likelihood estimation is used to find optimal $\widehat{\boldsymbol{\beta}}$ and variance of residuals σ^2 . The log-likelihood function is

$$\ln L(\boldsymbol{\beta}, \sigma) = \sum_{y_i<1} ln\left[\phi\left(\frac{y_i - \boldsymbol{x'_i\beta}}{\sigma}\right)/\sigma\right] + \sum_{y_i=1} ln\left[1 - \Phi\left(\frac{1 - \boldsymbol{x'_i\beta}}{\sigma}\right)\right] \tag{30}$$

where φ (·) and Φ (·) denote the probability and cumulative density functions for the standard normal distribution, respectively. Likelihood function is constructed by considering the probabilities of the dependent variable being below and above the boundary separately.

It's also sometimes used for LGD modelling two-tailed Tobit model. The two-tailed Tobit model uses a latent variable y* to model boundary cases such that $y_i * = \boldsymbol{\beta} * \boldsymbol{x_i} + \varepsilon_i$ where $y_i = min(1, max(0, y_i *))$. Assuming the distribution of the residuals conditional on **x** is normal, the following log-likelihood function is constructed for maximum likelihood estimation of **β** and variance of residuals σ^2 :

$$\ln L(\boldsymbol{\beta}, \sigma) = \sum_{y_i<1} ln\left[\phi\left(\frac{y_i - \boldsymbol{x'_i\beta}}{\sigma}\right)/\sigma\right] + \sum_{y_i=1} ln\left[1 - \Phi\left(\frac{1-\boldsymbol{x'_i\beta}}{\sigma}\right)\right] + \sum_{y_i=0} ln\left[\Phi\left(\frac{1-\boldsymbol{x'_i\beta}}{\sigma}\right)\right] \quad (31)$$

2.5 Beta Regression Model

LGD is essentially a ratio of the amount that the bank will not receive from the borrower to the total amount of debt, taking into account workout cost. This variable takes values in the range from zero to one with a large number of observations lying near zero and one. In this case, you can use the beta regression model, which is used for modelling when the explained variable takes values lying in a given interval

The first researchers who proposed to use of beta regression for estimation of Loss Given Default was Duan and Hwang[24]. They suppose that LGD has can be presented as:

$$LGD_i = \begin{cases} 0 & if\ X_i \in (-C_l, 0] \\ X_i & if\ X_i \in (0,1) \\ 1 & if\ X_i \in [1, C_u) \end{cases} \quad (32)$$

The beta distribution depends on two factors that determine its shape and therefore is very advantageous for modeling proportions. The beta density is defined as:

$$\pi(y;\ p, q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{p-1}(1-y)^{q-1} \quad (33)$$

With y from 0 to 1.

Where p>0, q>0, and $\Gamma(\cdot)$ is the gamma function.

We can express mean and variance of y as

$$E(y) = \frac{p}{p+q} \quad (34)$$

[24] DUAN, J.C., HWANG R.C. Predicting recovery rates at the time of corporate default, working paper, p. 19

And

$$var(y) = \frac{pq}{(p+q)^2(p+q+1)} \tag{35}$$

Cribari-Neto & Vasconcellos (2010) in their study estimate parameters by the maximum likelihood with adjustments and conclude that beta distribution is flexile distribution, which is cross functional in modelling and can be used in completely different areas, including economics.[25]

This distribution usually includes a significant number of zeros and / or ones. This can be a problem for estimation, since beta regression does not allow the presence of zeros and ones directly. In practice, this problem can be solved by transforming the dependent variable by adding / subtracting very small values to zero / one (as one thousandth) so that this transformation does not affect the whole distribution.

2.6 Zero-One Inflated Beta Models

From empirical data it is known that the distribution of the LGD often has a bimodal shape. After the default, two scenarios of further developments are most likely. In the first case, the debtor may regularly pay the recovery payments stipulated by the contract. In this case, the losses are relatively small and generally occurs due to administrative costs. In the second case, the debtor does not pay any payments or pays relatively small amounts that do not even cover the servicing costs. This usually leads to a loss amount increasing. This bimodality in the density of LGD distribution leads to difficulties in the development of possible evaluation models. Many approaches do not take this bimodal form into account.

As was mentioned a shortcoming of beta distribution is that it is not appropriate for modeling when data includes a lot of zeros and ones, which is typical for. This can be a problem when for estimation, since beta regression does not allow the presence of zeros and ones directly.

The discrete component is defined by a Bernoulli or a degenerate distribution at zero or at one. Zero-and-one inflated beta distributions is in fact mixture of a Bernoulli and beta distributions. It also can be one-inflated beta distributions which is respectively mixture

[25] CRIBARI-NETO, F., VASCONCELLOS K.L.P. (2010) Nearly Unbiased Maximum Likelihood Estimation for the Beta Distribution, p. 115.

of a beta and a degenerate distribution at zero. The proposed models are said to be inflated since the they allow for positive probability mass at some points (zero and/or one),in contrast with the beta. Proposed inflated beta regression models are the natural extensions models that were early introduced by Ferrari and Cribari-Neto (2004).[26]

It generates the response variable which is a mixture of Bernoulli and beta distributions, where true 0's and 1's, and the values between 0 and 1 are generated, respectively. The probability density function is

$$beinf(y;\alpha,\gamma,\mu,\phi)\begin{cases}\alpha(1-\gamma) & y=0\\ \alpha\gamma & y=1\\ (1-\alpha)f(y;\mu,\phi) & 0<y<1\end{cases} \quad (36)$$

Where $0<\alpha,\gamma,\mu<1$, and $\phi>0$. $f(y;\ \mu,\ \phi)$ is the probability density function for the beta distribution, parameterized in terms of its mean μ and precision ϕ:

$$f(y;\ \mu,\phi)=\frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)}y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1} \quad (37)$$

where $0<\mu<1$ and $\phi>0$.

α represents a mixture parameter that determines the extent to which the Bernoulli or beta component dominates to probability density function. Γ determines the probability that $y = 1$ if it comes from Bernoulli component. μ and ϕ are expected value and the precision for the beta component, which is usually parameterized in terms of p and q[27]

$$\mu=\frac{p}{p+q} \text{ and } \phi=p+q \quad (38)$$

2.7 Survival Analysis

When we have a group of subjects who stay in a certain state until a point in time at which a certain event occurs, we can use a survival analysis. Survival analysis is also called

[26] FERRARI, S.L.P., CRIBARI-NETO, F. Beta regression for modelling rates and proportions, p. 812

[27] OSPINA, R., FERRARI S. L. P. Inflated beta distributions. Statistical Papers, p. 121.

event analysis. The time to event is the time period from the subject's entry into the observations until his out.

This approach is widely used in completely different fields of activity. It was originally developed for medical research purposes, from which it got its name - the researchers analyzed level of survival after surgical interventions or the use of any treatment. Today, this approach is widely used in social sciences, engineering and economics, including credit management, for example for estimation of the time till recovery. Survival analysis is also called reliability or failure time analysis.

When considering the LGD in this approach, we try to estimate the time to recovery and the likelihood of making a payment in a certain period of time.

In the case of default loans, instead of patients, we consider individual cash payments that are in the collection process until they exit by a repayment.

Survival models consists of two components: $\lambda_0(t)$, which represents the underlying baseline hazard function. Baseline hazard function shows how the risk of event per time unit changes over time at baseline levels of covariates; and the effect parameters, describing how the hazard varies in response to explanatory covariates.

The hazard function can be representing by the formula:

$$\lambda(t) = \lim_{h \to 0+} \frac{P(t \leq T < t + h | t \leq T)}{h} \tag{39}$$

And specifies the instantaneous rate at which failures occur for items that are surviving at time *t*. The hazard function fully specifies the distribution of t and so determines both the density and survivor functions. From that follows:

$$\lambda(t) = \frac{f(t)}{S(t)} \tag{40}$$

Where S(t) is survival function, which can defined as:

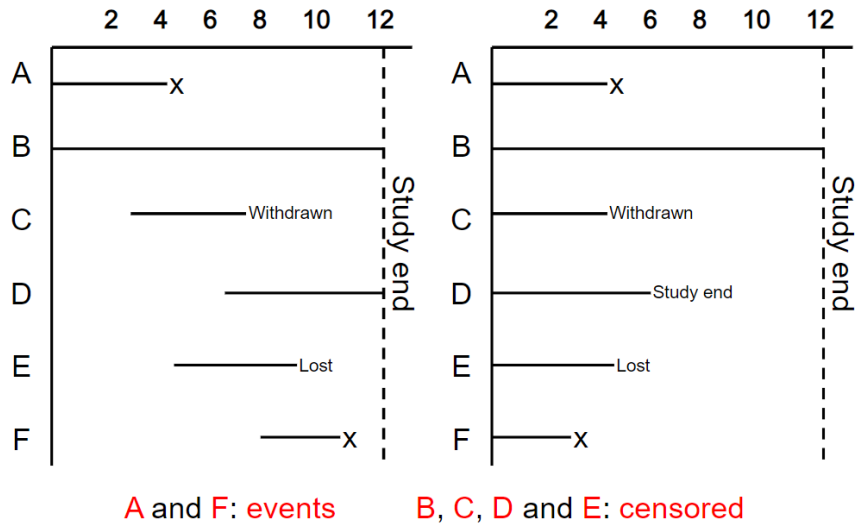
$$S(t) = exp\left[-\int_0^t \lambda(s)ds\right] = exp[-\Lambda(t)] \tag{41}$$

and show the probability of survival until time *t.*

From that $\Lambda = \int_0^t \lambda(s)ds$ is cumulative hazard function, so cumulative default rate can be defined as:

$$F(t) = 1 - exp[-\Lambda(t)] \quad (42)$$

In the case when we only know that the subjects are surviv until a certain point in time, but we have no other information about them, we can say that the observations are censored. Survival analysis is more flexible than other statistical methods and takes into consideration such indicators as censoring and time. Censoring can occur if the subject is out of observation before the end of the study. From a medical point of view, we can observe censored data, in the case when we only know the patient's status at a certain point in time. There are two basic options for censoring - left and right-side. If we have information about the state of the client before the observation beginning, we are talking about left-side censoring. If the patient is alive at a certain point in time – right – side censoring. In the case when we have only information that the event occurred in a certain period of time, this case is called interval censoring. In time to event data analysis all type of data are useful and bring some information that improves modeling accuracy, even censored.



Source: KARTSONAKI, C. Mini-Symposium: Medical Statistics: Survival analysis, p. 268

Figure 2. Censoring of survival data

In this approach, the subjects are not obliged to observe the same amount of time, that is, each subject can have their own entry and exit times and the duration of the observation,

and the analysis takes into account this, which is one of the main advantages of the survival analysis.

Another useful indicator is Kaplan-Meier estimation. Kaplan-Meier shows the proportion of patients who survived and continue to live for some time after an intervention. In various studies, the analysis of the effectiveness of intervention is evaluated by the ratio of patients who survived after it to the total number observed in this experiment. In the case of censored observations, for example, when a tested subject leaves the experiment until the end of observations for any reason, the Kaplan-Meier estimate is still effective, and it is also relatively simple, compared to the others, to calculate time-survival.

The Kaplan-Meier survival curve shows the probability of survival of the subject in each small time interval in period under review. The analysis assumes that all subjects have equal probabilities of survival, regardless of at what point in time they entered the study, in addition censored subjects also has the same survival probabilities as all other study participants. Thirdly, it supposed that the event occurs at the time specified. For better estimation of survival, it is proposed to conduct observations as often as possible.

Probability of survival at any particular time is calculated by the formula given below:

$$S_t = \frac{number\ of\ subject\ at\ begging - number\ of\ exits\ during\ time\ period}{number\ of\ subject\ at\ begging} \quad (43)$$

$$\hat{S}(t) = \prod_{i<t}\left(1 - \frac{d_i}{n_i}\right) \quad (44)$$

Subjects who did not survive or leave study for any reason are censored and these subjects aren't included in denominator. The overall probability of survival up to a specific point in time is obtained by multiplying all the probabilities of survival at all time intervals preceding a given moment (cumulative probability). The concept of conditional probability is also used. Despite the fact that in each interval we can have only a small number of observations, which affects the quality of the calculations, the overall probability of survival at each point is more accurate.

Consider estimating the cumulative hazard $\Lambda(t)$.

For estimation of cumulative hazard rate function we can use Nelson–Aalen estimator which is a nonparametric estimator. We can use this estimator for all type of data including censored. Nelson–Aalen estimator:

$$\hat{\lambda}(t_i) = \sum_{j=1}^{i} \frac{d_j}{n_j} \tag{45}$$

As we can see from the formula, in this case, the hazard rate is obtained from the ratio of deaths to the number of all subjects at each particular time. From here we can get a cumulative value by summing all the obtained values up to a certain point in time. We can interpret this rate as the expected number of deaths in time interval from the beginning of observation to certain moment per unit at risk.

The variance of $\hat{\lambda}(t_i)$ can be approximated by $var(-log\,\hat{S}(t))$.

Breslow (1972)[28] suggested estimating the survival function as:

$$\hat{S}(t) = exp\left(-\hat{\Lambda}(t_i)\right) \tag{46}$$

$(-\hat{\Lambda}(t_i))$ represents he Nelson-Aalen estimator of the integrated hazard. The Breslow estimator and the Kaplan-Meier estimator are asymptotically equivalent. In addition, in cases where the sample contains a relatively small number of deaths in relation to the whole number of subjects, these variables can be almost equal.

2.7.1 The Cox model

A frequently used regression model for analyzing survival data is the Cox proportional hazards regression model, sometimes also called semi-parametric. It takes its name because, when analyzed, it is able to estimate the relationship between the hazard rate and the explanatory variables without any assumptions about the form of the baseline hazard function

The Cox regression model is a statistical theory of counting processes that unifies and extends nonparametric censored survival analysis. The approach integrates the benefits of nonparametric and parametric approaches to statistical inferences.[29]

[28] BRESLOW L., BELLOC N.B. Relationship of physical health status and health practices. p. 421.
[29] COX, D.R.; OAKES, D. Analysis of survival data. p. 37

Let $X_i = \{X_{i1}, \dots X_{ip}\}$ be the realized values of the covariates for subject *i*. The hazard function for the Cox proportional hazards model has the form:

$$\lambda(t|X_i) = \lambda_0(t)\exp(\beta_i X_{i1} + \cdots + \beta_m X_{im}) = \lambda_0(t)\exp(X'\beta) \tag{47}$$

Where $\exp(X_i * \beta)$ determines like risk level.

The baseline hazard is a step function estimated on a discrete set of points where exits or censoring take place.[30] The corresponding survival function is in the form:

$$S(t, X') = exp\left(-\int_0^t \lambda_0(s) exp(X'\boldsymbol{\beta})\right) = S_0(t)^{exp(X'\boldsymbol{\beta})} \tag{48}$$

Where $S_0(t) = exp\left(-\int_0^t \lambda_0(s)\, ds\right)$ (49)

Xi denotes a covariate matrix for subject i. One or more of the variables in this matrix may vary over time. Hazard rates in the Cox model for two observations are proportional. Under assumptions of the model this proportionality remains unchanged over time.[31]

The coefficient vector **β** is estimated using the partial likelihood: if an object i, with covariate**s** $\mathbf{x}_i$ exits at time t_i, if we assume that there is only one exit at that time, and if Ai is the set of objects alive at time t_i, then the partial likelihood that just $i \in A_i$ exits is:

$$L_i(\boldsymbol{\beta}) = \frac{\lambda(t_i, x_i)}{\sum_{j \in Ai} \lambda(t_i, x_i)} = \frac{exp(-\boldsymbol{x_i}'\beta)}{\sum_{j \in Ai} exp(-\boldsymbol{x_i}'\beta)} \tag{50}$$

The coefficients β are then obtained by maximizing $ln\, L = \sum_{i=1}^{K} lnL_i$.

Given **β,** the baseline hazard function values are estimated separately for each of the time intervals, where the function assumed to be picewise constant maximizing the likelihood function:

$$L_t = \prod_{i=1}^{n} \prod_{t=1}^{T} [\lambda_0(t) exp(\boldsymbol{x'}_i\boldsymbol{\beta})]^{dN_i(t)}\ exp\left[-\int_0^T \lambda_0(u) exp(\boldsymbol{x}_i(\boldsymbol{u})'\boldsymbol{\beta}) Y_i(u) du\right] \tag{51}$$

[30] KALBFLEISCH, J. D., PRENTICE, R. L. The Statistical Analysis of Failure Time Data. p. 117
[31] COX, D.R. Regression Models and Life-Tables. p. 194.

Here $dN_i(t)$ is an index that indicates that the expected event occurred with the subject i at the time interval (t-1, t], and $Y_i(t)$ indicates of the fact that at the moment t-1 event still not occur.

If we differentiate log L with respect to the baseline function, for a fixed time t, we can get

$$\frac{\partial}{\partial \lambda_0(t)} \log L = \sum_{i=1}^{n} dN_i(t) - \sum_{i=1}^{n} \int_0^T \lambda_0(t) exp(\boldsymbol{x}_i(\boldsymbol{t})' \boldsymbol{\beta}) Y_i(t) \quad (52)$$

where we have used that λ0(t) is piecewise constant. Therefore, the maximum likelihood estimator for a baseline function has the form:

$$\widehat{\lambda_0}(t) = \frac{\sum_{i=1}^{n} dN_i(t)}{\sum_{i=1}^{n} exp(\boldsymbol{x'}_i \boldsymbol{\beta}) Y_i(t)} \quad (53)$$

As we can see, there is a number of different methodologies that are appropriate for Loss Given Default estimation. The second chapter represents part of them estimation frequently used as well as insufficiently studied in the existing literature. In total, seven different techniques were considered, some of which are parametric and the rest is non-parametric. Both research and practical experience today cannot reliably say which methods provide better estimations of parameters. Non-parametric methods, in contrast to parametric, represent a relatively new area that has to be studied in detail. The strength of parametric methods is their interpretability; however, they have worse prognostic abilities compared to non-parametric models that do not assume a specific distribution for LGD. In order to compare the obtained results with each other, this section also considers various approaches to measuring the model's fit, discusses the strengths and weaknesses of each approach.

3 Empirical model comparison

3.1 Data collection and structure of dataset

For estimation of model's parameters were used real loan data from public database "Bondora Capital" where public loan reports for different European country are presented. The data includes a rich set of variables collected at time of application and during the time of loan repayment. For data homogeneity only one country - Estonia was chosen.

The main database was obtained by combination of 3 databases:

- base with general loan data;
- historic payment base;
- secondary market transaction history.

Final dataset contains 126 variables and 29779 observations. The dataset provides with a great variety of information on borrower characteristics, contractual characteristics, additional data on defaulted loans. The richest source of predictor variables provides the information which was obtained at time of application for credit along with the credit bureau score collected by the bank i.e. the basic information about borrower: age, gender, marital status, working experience, income etc. Some credit history: value of previous borrower's loans, number of previous loans, number and amount of existing liabilities, how many times the borrower had repaid early. Loan's information: ap plied amount, purpose of the loan, credit duration, date when the loan was issued and maturity date, default date. Another potentially important group of information indicators is personal changes in circumstances over time. But the problem with this category is that it is difficult for a lender to extract data, in most of the cases, the bank cannot know exactly which reasons caused the client difficulty in making payments.

Historic payments base includes all received monthly payments in euro, categorized by repayment type (principal or interest repayment). Multiple payments for the same client were aggregated into a one single payment. Workout costs are also included into calculation, because neglecting the workout costs leads to underestimate the loss given default.

Data set contains 810 defaulted unsecured retail loans. In total, contracts have defaulted from 16.06.2009 till 01.11.2013. Historic payment base contains recovery payments from clients 20.10.2009 till 06.04.2018. Maximum recovery process is 60 months for all of the loans. The amount average exposure at default is about 390 EUR, maximum amount is slightly higher than 9900 EUR. The categorical variables have been split into several dummy variables. During recovery process loans can be fully repaid, written-off or sold.

All of the payments were discounted to reflect the time value of money and corresponding risk premium. Risk-free rate with risk premium were used for these purposes.

The bank's definition of default was applied, which is consistent with the regulatory framework; that is a contract is classified as defaulted if a borrower has become insolvent or is overdue with his payments on the underlying contract. For the defaulted contracts LGD on a contractual level were calculated.

LGD and, respectively, RR, can take values in the open interval from 0 to 1. There are cases when the borrower can pay amount larger than initial exposure at default, in this case, the LGD will be negative. In case, when the debtor does not pay anything and moreover the bank has additional workout costs, there will be LGD higher than 1, however, we will exclude such cases from consideration and take variable values from 0 to 1. LGD has characteristics of a bimodal distribution, since the most frequently observed values are extreme - zero and one, what is an absolute absence of payments or alternatively repayment of the full amount.

Table 1. Descriptive statistics of recovery rate

Statistic	Value
Number	810
Min	0
Max	1
Mean	0,460
Median	0,568
Standard deviation	0,487

Source: own calculations

Distribution of recovery rate shows Figure 4. Average value of recovery rate is 0,46 with 29% of observations in data that have recovery rate equal zero.

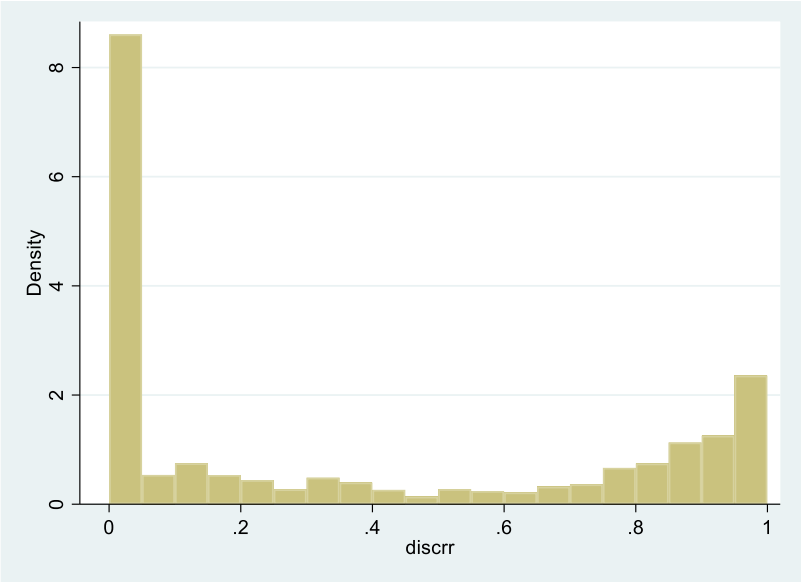


Figure 3. Distribution of recovery rate
Source: own illustration

3.2 Empirical Results

In this chapter, the obtained outcomes are presented.

Seven modelling algorithms are considered: linear and logistic regressions, Tobit model, regression tree model, Beta regression model, Zero-One Inflated Beta model and survival analysis represented by the Cox and modified Cox regressions.

Models was calibrated on randomly divided training sets of 70% and validation of their performance was provided on the remaining 30% of the original dataset. Out-of-sample results are presented in this part.

All presented calculations were performed using software Stata and RStudio. Variables for models were chose by stepwise procedure with small corrections in several cases.

Linear regression

Final model includes 4 variables – initial duration of the loan, gender of borrower (0 – male, 1 – female), marital status of borrower (1 – married , 0 – other) and of higher education (1 – yes, 0 – no). As we can see from the Table 2, all of the P – values are less

than 0,1 that means all variables are significant on 10% significance level. Variable loan duration has negative sign, that means hihger duration leads to smaller potenional loss given default.

Table 2. Regression outputs

Variable	Coefficient	Standard error	P-value
Loan duration	-0,013	0,003	0,000
Gender	-0,137	0,054	0,012
Marital status	0,164	0,057	0,004
Higher education	0,23	0,12	0,051

Source: own calculations

It's interesting fact, that women tend to pay higher proportion of exposure at default in comparison with men.

Table 3. Model performance quality

R^2
0,09870717

Source: own calculations

Logistic regression

For logistic regression it is necessary to select the threshold to assign values 0 and 1 to LGD variable. Model has the same independent variables as linear regression has and coefficient's signs are also the same.

Table 4. Regression outputs

Variable	Coefficient	Standard error	P-value
Loan duration	-0,069	0,019	0,000
Gender	-0,721	0,293	0,014
Marital status	0,656	0,312	0,035
Higher education	1,163	0,714	0,093

Source: own calculations

This model has higher R^2 than previous.

Table 5. Model performance quality

R^2
0,15063278

Source: own calculations

Censored Linear Regression (Tobit) Model

Since LGD has censored distribution with a with values in the range from zero to one with a large number of observations lying near zero and one we can provide estimation with Tobit model. Tobit takes into consideration this censoring of the dependent variable through truncation. The Tobit model uses a latent variable y* to model boundary cases such that $y_i^* = \beta \cdot x_i + \varepsilon_i$ where $y_i = max(y_i^*, 1)$ for right-sided truncation. A likelihood function can be constructed, assuming the distribution of the residuals conditional on x is normal.

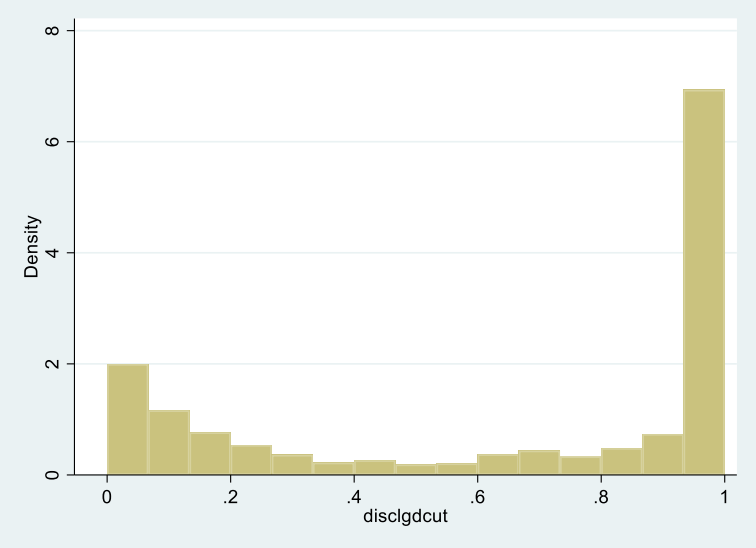


Figure 4. Loss given default distribution
Source: own illustration

As was mentioned, Tobit model is applicable in only those situations, where latent variable can, in principle, be higher than 1 or lower than 0, but these values are not observed because of censoring. Where these values are a consequence of individual decision, these decisions should be modeled appropriately and the Tobit model should not be used mechanically.[32]

[32] MADDALA, G.S. Limited-Dependent and Qualitative Variables in Econometrics. p. 247.

Table 6. Tobit regression outputs

Variable	Coefficient	Standard error	P-value
Loan duration	-0,016	0,003	0,000
Gender	-0,161	0,067	0,017
Marital status	0,217	0,072	0,003
Higher education	0,359	0,159	0,025

Source: own calculations

Final Tobit model includes the same 4 variables as prvious models, but has slightly different coefficient, however signs are the same so overall logic of estimation is kept.

Table 7. Model performance quality

R^2
0,02433768

Source: own calculations

Coefficient of determination is very low, what is unexpected result because Tobit model usually provide sufficiently good fit.

As long as LGD usually follows a beta distribution we can estimate it using beta regression. There is a one limitation with basic beta-regression in Stata: estimation procedure ignores 0 and 1. There are several ways to include it in analysis.

Beta regression model

First way is to use beta regression. To estimate data, we can transform dependent variable to "push" 0s and 1s a tiny bit inwards (0,0001 and 0,9999) what should not affect the result, but allow not exclude observations from the analysis.

Table 8. Regression outputs

Variable	Coefficient	Standard error	P-value
Loan duration	-0,042	0,0112	0,000
Gender	-0,277	0,177	0,017
Marital status	0,552	0,185	0,003
Higher education	0,764	0,384	0,047

Source: own calculations

The regression has the same explanation variables as the previous regressions and coefficients has the same signs. This model has one of the best performances.

Table 9. Model performance quality

R^2
0,14071137

Source: own calculations

Zero-one inflated Beta

Second way to include zeros and ones is zero one-inflated beta model. This model is for situations where you believe that the decisions for proportions of 0 and/or 1 are governed by a different process as the other proportions. For this model different explanation variables were chosen for LGD values zero, one and for values between.

In this case, in fact, estimation goes at the same time parallel by three regressions. One of them separates ones from all other values and provide estimations, the second does the same with zeros and the third estimates separately values in the interval between zero and one not including the extreme values themselves.

Table 10. Regression outputs

Variable	Coefficient	Standard error	P-value
Proportion			
Loan duration	-0,0264	0,009	0,007
Marital status	0,622	0,169	0,000
Existing liabilities	-0,105	0,05	0,037
Const	0,855	0,163	0,000
One-inflate			
Loan duration	-0,044	0,012	0,000
Higher education	2,537	0,788	0,001
Employment duration	-0,048	0,284	0,090
Const	0,064	0,208	0,758
Zero-inflate			
Gender	0,449	0,217	0,039
Age	-0,017	0,01	0,094
Employment duration	-0,717	0,283	0,013
Const	0,205	0,356	0,910

Source: own calculations

The logic of this approach is to assume that there are some qualitative differences in the factors that encourage people to pay the full amount or pay less. It is also possible to say

that customers who did not pay nothing at all have different characteristics, compared with those who paid at least something or amount in full.

The first two options are estimated using logistic regression, the third through a beta regression. All of the «processes» include the same numbers of variables, but explanatory variables are different.

But the model has a low explanatory power:

Table 11. Model performance quality

R^2
0,06441946

Source: own calculations

Survival analysis

The vertical lengths between horizontal lines show the change in cumulative probability of survival, while horizontals represent duration of survival. The cumulative probability of surviving a given time is seen on the Y-axis. The Kaplan-Meier curve is rather step-wise estimate but not the smooth functions, because of its non-continuous nature; so, it can be difficult to estimate survival in a particular point of time, while the probability is given on a specific length of time. For clarity probability of surviving 10 months is around 50%, probability of surviving 30 months is about 25%, the steepness of the curve is determined by the survival durations (length of horizontal lines).

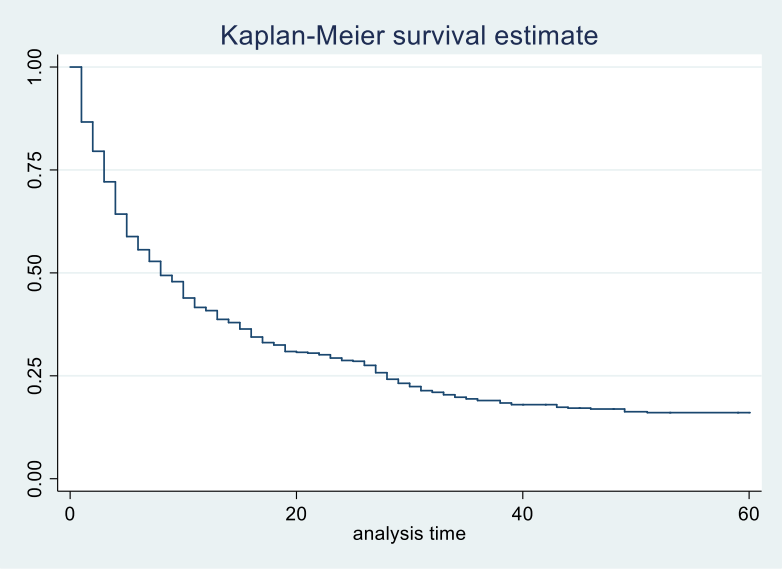


Figure 5. Kaplan-Meier curve
Source: own illustration

Cox model divides the exposure at default into individual monetary units and follows their survival in time. Units that was not repaid till certain time are considered as survived at this point in time.

Table 12. Cox regression model outputs

Variable	Coefficient	Standard error	P-value
Loan duration	0,056	0,003	0,000
Number of previous loans	0,07	0,011	0,000
Higher education	-0,93	0,180	0,000
Employment duration	0,19	0,079	0,017

Source: own calculations

From the Table 12 we can see that this regression includes partially other explanatory variables, which can be explained by the fact that the regression is based on a different logic and has a different dependent variable. In addition, the signs of coefficients differ from the previous models.

Table 13. Model performance quality

R^2
0,0269

Source: own calculations

Modified Cox model

In modified Cox model time till full recovery was estimated i.e. in that case event of interest is 100% recovery repayment. In contrast with classic Cox model here we do not base our estimate separate monthly payments, but the whole amount of debt.

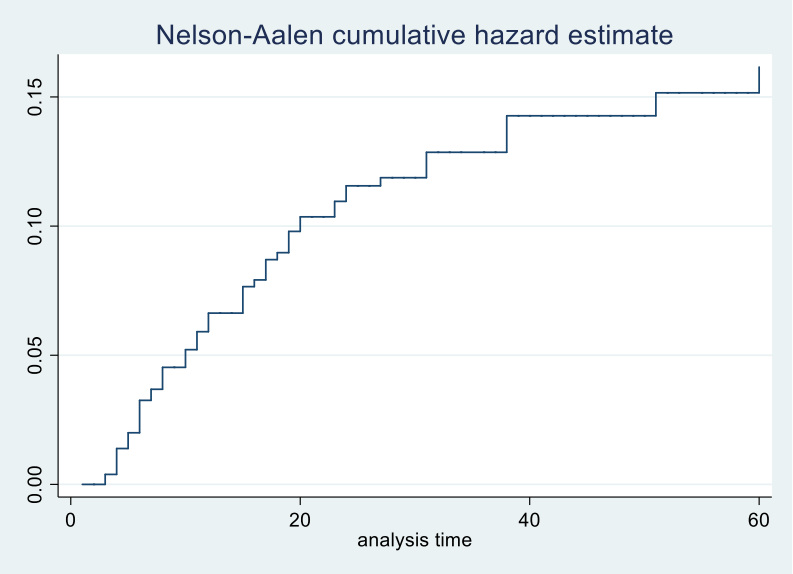


Figure 6. Nelson-Aalen curve
Source: own calculations

The Nelson–Aalen estimator is a nonparametric estimator which may be used to estimate the cumulative hazard rate.

Table 14. Modified Cox regression model outputs

Variable	Coefficient	Standard error	P-value
Gender	0,542	0,2665	0,042
Age	0,036	0,0134	0,006
Liabilities to income	0,273	0,086	0,002
Amount of previous loans	-0,002	0,0006	0,001
Employment duration	-1,796	0,594	0,003

Source: own calculations

Table 15. Model performance quality

R^2
0,0735

Source: own calculations

Regression trees

These algorithms produce trees using "if-then" conditions to divide the data sequentially in order to reduce its inhomogeneity. Final subsamples that were determined by the process then are averaged in accordance with their loss given default value. Construction

of the regression tree is based on the dividing of original sample with certain threshold values.

When additional splits yield doesn't increase explanation power or a minimum of instances per subset is reached, then the split is optimal in terms of maximized gain ratio is performed. Every partition result is in a node.

A fully developed tree may have a terminal node for each observation, but in this case, tree can be overfitted and useless for out-of-sample prediction. In contrast, an extremely flat tree may have low explanatory power. However, correct stop criterion to the trees ensures that overfitting or underfitting would be mostly avoided.

In this paper, the mechanism is implemented in such a way that the fulfillment of the condition in the node corresponds to further movement to the left, the failure to fulfill the condition to the right

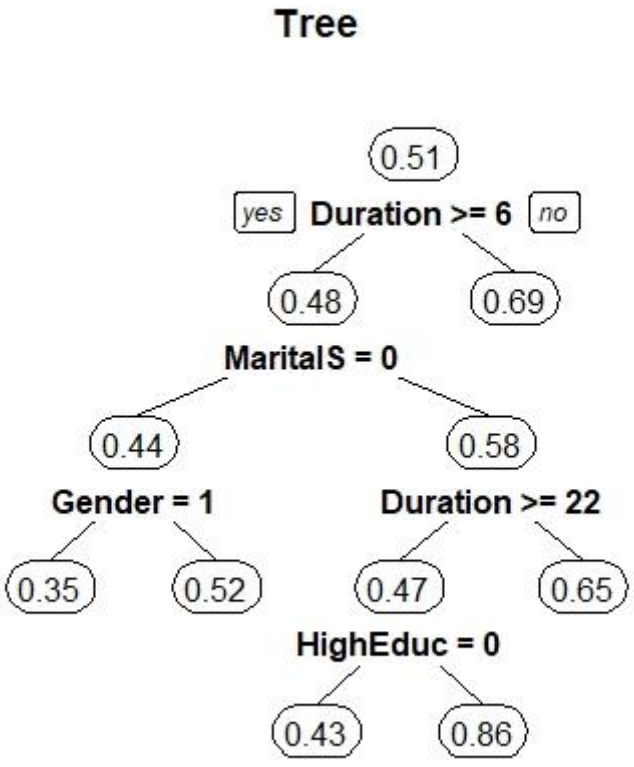


Figure 7. Regression tree estimates
Source: own illustration

As we can se, four variables were choosen as explanatory for LGD in this model: loan duration, marital status, gender and higher education.

We are going to want to make sure that we aren't overfitting the data, so let's prune the tree. An easy way to do this is by looking at the complexity parameter, which is basically the "cost", or decrease in performance, of adding another split, and choose the tree size that minimizes our cross-validated error.

Following graph shows the optimal size of regression tree.

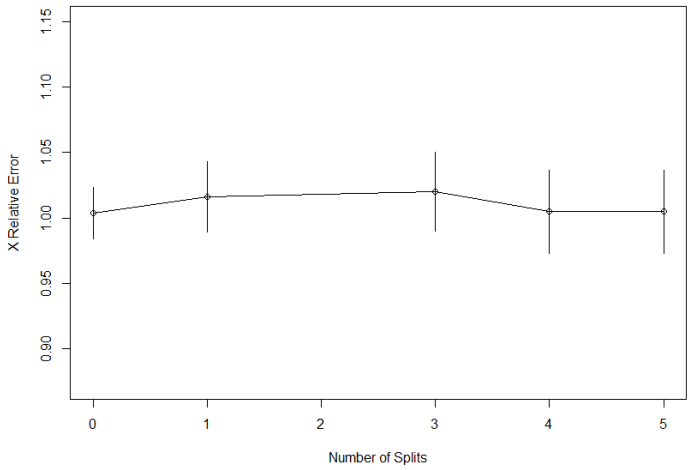


Figure 8. Optimal size of regression tree
Source: own illustration

In our case optimal size of a tree is four or five nodes.

Table 16. Model performance quality

R^2
0,0933

Results are reported in Table 17. These results demonstrate that the best performance power have logistic regression and Beta regression model.

Table 17. Comparison of results

Method	R-squared
Linear regression	0,0987
Logistic regression	0,1506
Tobit regression	0,0243
Beta regression	0,1407
Inflated beta regression	0,0644
Cox model	0,0269
Modified Cox model	0,0735
Regression tree	0,0933

Source: own calculation

Conclusion

The aim of this paper is to compare different estimation's methodologies of one of the most important Basel indicators – loss given default and to find the more appropriate model.

The first chapter of this work briefly describes the main provisions of the Basel II and the most important its indicators. The history of the Basel Accord development, basic requirements and approaches to estimation is briefly described in the first part of the chapter. The second part articulates the concept of loss given default in detail. The First Basel Accord was introduced in 1980s and for today contains capital risk, operational risk and market risk requirements and recommendations. Main purpose of the Basel is to provide requirements in regards to minimum capital of financial institutions to meet their obligations and absorb unexpected losses. LGD is one of the most important parameters used in the calculations of the Basel risk weight function and the regulatory capital for credit risk. When estimating LGD, it is necessary to accurately quantify the level of potential recoveries obtained after the default, since it is the basis for the calculation of regulatory capital, which is the foundation of the bank's stability. Currently, Basel offers several options for modelling the main risk parameters, the preferred of which is the one, where the bank internally calculate parameters since this method is more reliable and more accurately reflects the existing risks of a particular bank. At the same time, the implementation of this method is associated with certain difficulties. Currently, there is a broad theoretical and practical base of recommendations for calculating the probability of default, while for the loss of default there is no structured guide to the selection and comparison of models.

So, the second chapter represents several methods of Loss Given Default estimation frequently used as well as insufficiently studied in the existing literature. All the methods used are described in detail. In total, seven different techniques were used in this work, some of which are parametric and the rest is non-parametric. Both research and practical experience today cannot reliably say which methods provide better estimations of parameters. Non-parametric methods, in contrast to parametric, represent a relatively new area that has to be studied in detail. The strength of parametric methods is their interpretability; however, they have worse prognostic abilities compared to non-parametric models that do not assume a specific distribution for LGD. In order to compare

the obtained results with each other, this section also considers various approaches to measuring the model's fit, discusses the strengths and weaknesses of each approach. In particular, such methods as Mean Squared Error, Root Mean Square Error and Mean Absolute Error were described besides the commonly used weighted R-squared.

The last chapter presents the results of empirical evaluation conducted on real data. Analysis were provided on a loan data from public database "Bondora Capital" where public loan reports for different European country is presented. The data includes a rich set of variables collected at time of application and during the time of loan repayment. the database contains a fairly large number of indicators for which data are incomplete or missing, but despite this, it was possible to select enough suitable variables for analysis.

Different approaches to Loss Given Default modelling were considered: linear regression, logistic, Tobit model, Beta model, zero-one inflated beta model, survival analysis, and regression trees. These models were applied to the real data in order to compare their performance.

For comparison of model fit standard goodness of fit measure represented by EAD weighted R-squared was used. Results show that in particular case of this paper the best prediction power has logistic regression model, slightly lower but almost the same result has Beta regression model.

Results show that the best performance provide logistic regression, which has the highest $R^2 = 0{,}1506$. This is not surprising, because of the high explanatory power and simplicity of using of logistic regression is often used in practice and in theoretical studies.

Slightly lower, but almost the same coeficient has Beta regression models. As was mentioned in theoretical part, regression and classification trees usually has lower quality of explanation and prognosis, our practice confirm this statement, the model has an average R^2 which is on par with standard linear regression and Zero-One Inflated Beta model.

Surprisingly low R^2 have a Cox regressions, both standard and modified. For a modified version, this may be due to a small number of observations - a relatively small number of client have fully repaid the loan after default.

One of the lowest R-squared coefficient has Tobit regresion, what is also unexpected result because Tobit model usually provides sufficiently good fit.

From all of the above, we can conclude that each of the models has its pros and cons and the quality of the estimation basically depends not only on the model chosen, but also on data used in a particular case, the selected variables and many other factors.

Because banking is inherently risky, the health of banks in large measure depends on their ability to manage risk and the associated exposure to losses. Banks should pay more attention to model development and validation because of importance of banking health for the overall stability of the financial system and economy.

References

1. ALTMAN, E. I.; KISHORE V.M., Almost Everything You Wanted to Know about Recoveries on Defaulted Bonds, *Financial Analysts Journal*. 1996, Vol. 52, No. 6. pp. 57-64

2. ASARNOW, E.; EDWARDS, D. Measuring loss on defaulted bank loans: a 24-year study, *Journal of Commercial Lending*. 1995, vol. 77, n.7, pp. 11-23.

3. BASTOS, J., Forecasting bank loans loss-given-default. *Journal of Banking & Finance*. 2010, Vol. 34, Issue 10, pp. 2510-2517

4. BCBS (2005): Basel Committee on Banking Supervision, Guidance on Paragraph 468 of the Framework Document. Basel, Basel Committee on Banking Supervision, 2005.

5. BCBS (2006): International Convergence of Capital Measurement and Capital Standards. A Revised Framework – Comprehensive Version. Basel, Basel Committee on Banking Supervision, 2006.

6. BRESLOW, L.; BELLOC, N.B. Relationship of Physical Health Status and Health Practices, *Preventive Medicine*. 1972, 1, pp. 409-421.

7. BRIEMAN, L.; FRIEDMAN, J. H.; OLSHEN R. A.,; STONE C. J. Classification and Regression Trees. Chapman & Hall/CRC. 1984.

8. CHALUPKA, R., KOPECSNI, J. Modeling Bank Loan LGD of Corporate and SME Segments: A Case Study. Finance a uver. *Czech Journal of Economics and Finance*. 2009, 59, no. 4

9. COX, D.R.; OAKES, D. Analysis of survival data. 1984, 212p.

10. COX, D.R. Regression Models and Life-Tables. *Journal of the Royal Statistical Society*. Series B (Methodological). 1972, Vol. 34, No. 2., pp. 187-220.

11. CRIBARI-NETO, F.; VASCONCELLOS, K.L.P. Nearly Unbiased Maximum Likelihood Estimation for the Beta Distribution, *Journal of Statistical Computation and Simulation.* Vol. 72, 2002 - Issue 2, pp. 107-118

12. CROOK, J.; BELLOTTI, T. Loss given default models incorporating macroeconomic variables for credit cards, *International Journal of Forecasting*. 2012, vol. 28, no. 1, pp. 171-182.

13. DUAN, J-Ch.; HWANG R-Ch. Predicting recovery rates at the time of corporate default, working paper, *National University of Singapore*. 2014.

14. FERRARI, S.L.P; CRIBARI-NETO, F. Beta regression for modelling rates and proportions. *Journal of Applied Statistics*. 2004, 31(7): 799-815

15. GORDY, M.B.; CAREY, M.J. Measuring Systematic Risk in Recoveries on Defaulted Debt I: Firm-Level Ultimate LGDs. 2004.

16. GREENE, W.H. Econometric Analysis. Upper Saddle River, NJ: Prentice Hall. 2003

17. GUPTON, G. M.; STEIN R.M. (2002). "LossCalc™: Model for Predicting Loss Given Default (LGD)." Moody's Investors Service.

18. GUPTON, G.M., Advancing Loss Given Default Prediction Models: How the Quiet Have Quickened, *Economic Notes, Banca Monte dei Paschi di Siena SpA*. 2005, vol. 34(2), pages 185-230, July.

19. HOSEMR. D.W.; LEMESHOW. S., Applied Survival Analysis: Regression Modeling of Time-to-Event Data, 2nd Edition.1976

20. HURLIN C.; LEYMARIE J.; PATIN A. Loss functions for Loss Given Default model comparison, *European Journal of Operational Research*, *Elsevier*. 2018, vol. 268(1), pages 348-360.

21. JACOBS M. An Empirical Study of the Returns on Defaulted Debt and the Discount Rate for Loss-Given-Default. 2009

22. JOHNSON, N. L.; KOTZ, S.; BALAKRISHAN N. Continuous Univariate Distributions, vol. 2, 2nd edn (New York: Wiley). 1995.

23. KALBFLEISCH J.D.; PRENTICE R. L. The Statistical Analysis of Failure Time Data. Hoboken, New York, Wiley, 2002.

24. KALOTAY E.A.; ALTMAN E.I. Intertemporal Forecasts of Defaulted Bond Recoveries and Portfolio Losses. *Review of Finance*. 2017, 21(1), pp. 433-463.

25. KARTSONAKI, C. Mini-Symposium: Medical Statistics: Survival analysis. *Diagnostic Histopathology*. Vol. 22, Issue 7, 2016, pp. 263-270

26. KAPLAN E. L.; MEIER P. Nonparametric Estimation from Incomplete Observations, *Journal of the American Statistical Association*. Vol. 53, No. 282, 1958, pp. 457- 481.

27. LI P., ZHANG X., ZHAO X. (2018) Modeling Loss Given Default (July 1, 2018). FDIC Center for Financial Research Paper No. 2018-03.

28. LOTERMAN, G.; BROWN, I.; MARTENS, D. Benchmarking regression algorithms for loss given default modeling, *International Journal of Forecasting*, vol. 28, 2012. pp 161–170.

29. MACLACHLAN I. Choosing the discount factor for estimating economic LGD in Altman E., Resti A. and Sironi A. (eds), Recovery risk. The next challenge in credit risk management, Risk Books, London. 2004.

30. MADDALA, G.S. Limited-Dependent and Qualitative Variables in Econometrics. Cambridge University Press. 1983. 416 p.

31. NAZEMI, A., FATEMI POUR F., HEIDENREICH, K., FABOZZI, F.J.. Fuzzy Decision Fusion Approach for Loss-Given-Default Modeling. European Journal of Operational Research, 262(2), 2017. pp. 780-791.

32. OSPINA R.; FERRARI, S. L. P. Inflated beta distributions. Statistical Papers, 51, 111–126. 2010.

33. OSPINA R.; FERRARI, S. L. P. A general class of zero-or-one inflated beta regression models. *Computational Statistics & Data Analysis.* Volume 56, 2012, pp. 1609-1623

34. PRIVARA S.; KOLMAN S.; WITZANY J., Recovery Rates in Consumer Lending: Empirical Evidence and the Model Comparison, *Bulletin of the Czech Econometric Society, The Czech Econometric Society*. 2014, vol. 21(32).

35. QI, M.; ZHAO X. Comparison of modeling methods for loss given default, *Journal of Banking and Finance.* 2011, 35 (11), pp. 2842-2855.

36. SOMERS M.; WHITTAKER J. Quantile regression for modelling distributions of profit and loss. *European Journal of Operational Research*. 2007, 183(3),1477-1487

37. TOBBACK, E., MARTENS, D., VAN GESTEL, T., BAESENS, B, Forecasting loss given default models: Impact of account characteristics and the macroeconomic state. *Journal of the Operational Research Society*, 65 (3), 2014. pp. 376–392.

38. WITZANY, J.; RYCHNOVSKY, M.; CHARAMZA, P. Survival Analysis in LGD Modeling. IES Working Paper. 2/2010.

39. WITZANY, J. Credit risk management - Pricing, Measurement, and Modeling. 1st ed. Swiss: Springer International Publishing, 2017. 250 p.

40. WITZANY, J. Unexpected Recovery Risk and LGD Discount Rate Determination, *European Financial and Accounting Journal*. 2009, vol. 4, no. 1, pp. 61-84.

41. YANG B.H.; TKACHENKO M. Modeling of EAD and LGD: Empirical Approaches and Technical Implementation. 2012.

42. YASKIR, O.; YASKIR, Y. Loss given default modeling: Comparative analysis. *Journal of Risk Model Validation*, v.7, No.1, 2013.

43. ZHANG J.; LYN C. T. Comparison of linear regression and survival analysis using single and mixture distribution approaches in modelling LGD. [in special issue: Special Section 1: The Predictability of Financial Markets. Special Section 2: Credit Risk Modelling and Forecasting] *International Journal of Forecasting*. 2012, 28 (1). pp. 204-215.