Lappeenranta-Lahti University of Technology LUT
School of Business and Management
Business Administration

Tuomas Savolainen

ESTIMATING LOSS GIVEN DEFAULT FOR RETAIL CREDIT PORTFOLIO

Examiners: Professor Eero Pätäri
Associate professor Sheraz Ahmed

ABSTRACT

Author:	Tuomas Savolainen
Title:	Estimating loss given default for retail credit portfolio
Faculty:	LUT School of Business and Management
Master's program:	Strategic Finance and Analytics
Year:	2021
Master's thesis:	Lappeenranta-Lahti University of Technology LUT 68 pages, 9 figures, 8 tables and 3 appendices
Examiners:	Professor Eero Pätäri Associate professor Sheraz Ahmed
Keywords:	loss given default, internal ratings-based approach, retail credit risk management, banking, minimum capital requirement

Financial institutions have the option to internally estimate LGD and PD in determining risk weight of loans, instead of the standard approach. This thesis contributes to the existing literature by employing five different methodologies in estimating LGD for retail credit portfolio. Earlier research has applied multiple techniques, but no consensus exists for the best practice of LGD modeling.

The research composes of forecasting LGD and comparing model estimates. The data set and models are built to comply with the effective Basel framework and EU legislation. The models in use are linear regression, generalized linear regression (gamma regression), beta regression, random forest, and support vector regression. Prediction performance was evaluated from perspective of estimation accuracy and discriminatory power with five different metrics in use. Overall ranking was formed by ranking models from the best performance to the worst with individual metrics and aggregating the rankings. The best performing model was generalized linear regression with gamma distribution in use. This thesis concludes that LGD models are case sensitive as they are built on the need of individual institutions and their specific risk characteristics hence the lack of single best methodology.

TIIVISTELMÄ

Tekijä:	Tuomas Savolainen
Tutkielman nimi:	Tappio-osuuden estimointi vähittäisluotoille
Tiedekunta:	Kauppatieteellinen tiedekunta
Pääaine:	Strateginen rahoitus ja analytiikka
Vuosi:	2021
Pro Gradu-tutkielma:	Lappeenrannan-Lahden teknillinen yliopisto LUT 68 sivua, 9 kuviota, 8 taulukkoa, 3 liite
Tarkastajat:	Professori Eero Pätäri Tutkijaopettaja Sheraz Ahmed
Avainsanat:	tappio-osuus, sisäisen luottoluokituksen menetelmä, luottoriskin hallinta, pankkitoiminta, minimi pääomavaade

Rahoituslaitoksilla on mahdollisuus käyttää sisäisen luottoluokituksen menetelmää estimoidakseen LGD- ja PD-arvoja, joita hyödynnetään riskipainolaskennassa standardimallin sijasta. Tämä pro gradu tutkielma täydentää olemassa olevaa kirjallisuutta käyttämällä viittä eri metodologiaa LGD:n estimointiin vähittäisluotoille. Aiempi kirjallisuus on tutkinut useaa eri tekniikkaa, mutta yhtenevää näkemystä parhaasta metodologiasta ei ole syntynyt.

Tämä tutkimus koostuu LGD-estimaattien muodostamisesta valituilla metodeilla ja niiden ennusteiden arvioinnista. Käytetty aineisto ja mallit täyttävät voimassa olevat Basel kehikon ja EU:n sääntelyvaatimukset soveltuvin osin. Tutkimuksessa käytettävät mallit ovat lineaarinen regressio, yleistetty lineaarinen regressio (gamma regressio), beeta regressio, satunnainen metsä ja tukivektori regressio. Ennusteiden suorituskykyä arvioidaan tarkkuuden ja luokittelukyvyn mukaan viidellä mittarilla. Mallien paremmuusjärjestys määritetään ensin yksittäisten metriikoiden mukaan ja lopulta järjestysten summana. Parhaiten suoriutunut malli oli yleistetty lineaarinen regressio gamma jakaumaa soveltaen. Tämä pro gradu tutkielma vahvistaa LGD-mallien tapauskohtaisuuden, sillä ne rakennetaan vastaamaan yksittäisten instituutioiden riskiprofiiliin ja tarpeisiin, minkä takia yksi metodologia ei ole noussut toisten yläpuolelle.

ACKNOWLEDGEMENTS

I want to thank the LUT-university for providing a platform for growth as person and as a business professional throughout my time in Lappeenranta. Atmosphere in LUT campus is very inspiring especially in the library.

Thanks to the incredible colleagues and friends for the memories we shared studying and in other activities. Praise to all the volunteer people at Business student guild Enklaavi and the Student union of LUT University for networking possibilities.

A warm thank you to my relatives for supporting me while writing this thesis and providing help along the whole journey. Finally, special thanks to Peppi for providing incredible cheering and endless support even in the most difficult of times.

In Lappeenranta, 23rd August 2021

Tuomas Savolainen

Table of contents

Appendices

List of Tables

List of Figures

List of Abbreviations

BCBS	Basel Committee on Banking Supervision
BIS	Bank of International Settlements
CRD	Capital requirements directive
CRR	Capital requirements regulation
EAD	Exposure at default
ELBE	Expected loss best estimate
IRB	Internal ratings-based
LGD	Loss given default
MAE	Mean absolute error
MSE	Mean squared error
PD	Probability of default
RDS	Reference data set
RF	Random forest
RR	Recovery rate
RWA	Risk-weighted assets
SME	Small to medium size enterprises
VAR	Value-at-risk
WOE	Weight of evidence

1. Introduction

Credit is the main tool to build economy, allocate capital efficiently, and generate wealth. (Koulafetis 2017) Its headliner role in financial system has raised the need to form concise rules in order to control and enhance the systems resilience. In recent decades international banking regulation has been formed under Basel Committee for Banking Supervision (BCBS). This attempt to equalize global playing field has produced a broad regulatory framework for financial institutions and it has specifically targeted international banks.

1.1. *Background and Motivation of the Study*

BCBS sets regulatory standards on calculation of regulatory capital for banks. Institutions have two options: standard or internal model to fulfill this requirement. One major difference in the methods is the risk weight determination when calculating the risk weighted assets (RWA). In standard models the RWA are given by supervisor. In the internal ratings-based (IRB) approach institutions estimate the RWA individually. IRB RWA formula consists of probability of default (PD), exposure at default (EAD), and loss given default (LGD). (BIS 2008, Regulation (EU) No 575/2013) Among these parameters, PD is the most researched and best business practices are somewhat formed. In contrast LGD and EAD have had less research although they are equally important parameters. (Huajian & Tkachenko 2012) Former LGD studies have been targeting company credit while retail studies are scarce. One reason for this is that company information, for example financial reports, are more easily available than retail information which is usually confidential. (Leow 2012) For these reasons this thesis aims to bring new insights on field of retail LGD modeling.

LGD has been a subject of research in many studies but there seems to not be a clear and concise consensus on the topic. EU (No 575/2013, art. 4 point 55) regulation defines LGD as “ratio of the loss on an exposure due to the default of a counterparty to the amount outstanding at default”. Regulatory definition is precise while estimation methods have been versatile. For example, Schuermann (2004) uses a model based on loan characteristics derived from the debt collection process. On the other hand, Zhang & Thomas (2012) estimated LGD indirectly from loan recovery rate (RR). This paper will analyze retail credit portfolio with different lines of credit which differs from past research. Objective is to bring new insights on characteristics of retail credit risk and how to model it accurately. Resulting LGD model can be used as baseline model that complies with IRB regulation. Another motivation for this paper is to compare different modeling techniques in finding the best performing method for retail credit portfolio.

Weight of evidence (WOE) method is known widely from credit scorecards and ratings. Credit scoring is a process to rate borrower's according to individual risks to default. (Yap et al. 2011) WOE can be used for optimal selection of variables and binning (sometimes called bucketing or discretizing) of both numerical and categorical variables. WOE transformation has been found useful in other risk related models, but it is not widely used in context of Basel IRB regulation. (Huajian & Tkachenko 2012) This study will apply WOE transformation to test the performance and asses technical implementation.

1.2. *Purpose of the Study*

Purpose of this study is to research LGD modeling and identify best modeling technique for retail credit portfolio. Since one concise standard for LGD modeling has not been achieved this study will examine several ways to do it. This research will target performing exposures so that the proposed model could be applied to estimate LGDs of healthy loans. In context of financial institutions and internal methods understanding regulation is a key concept. This study will discuss international and local level regulation and compliance standards required for internal estimation of LGD. Main research question and sub-questions are as follows:

Research questions:

How to forecast LGD in retail credit portfolio?

What is the best method to forecast LGD?

1.3. *Methodology and focus of the study*

Main focus of this study is on credit risk, more specifically in quantifying and measuring it. In Figure 1 we can see the theoretical framework where credit risk is one part of risk management. Risk management is studied from the perspective of a bank. Other major risks to which banks are exposed are market and operational risks (Win 2018, 744). This thesis focuses on credit risk from two perspectives: capital adequacy and debtor specific risks. Capital adequacy is important for credit risk quantification as banks need to hold substantial amounts of own capital compared to exposures. One part of this thesis focuses on regulation. Debtor specific risks are studied in order to form a good and efficient estimation model for individual loan LGDs. This is done similarly to how the probability of default is estimated.

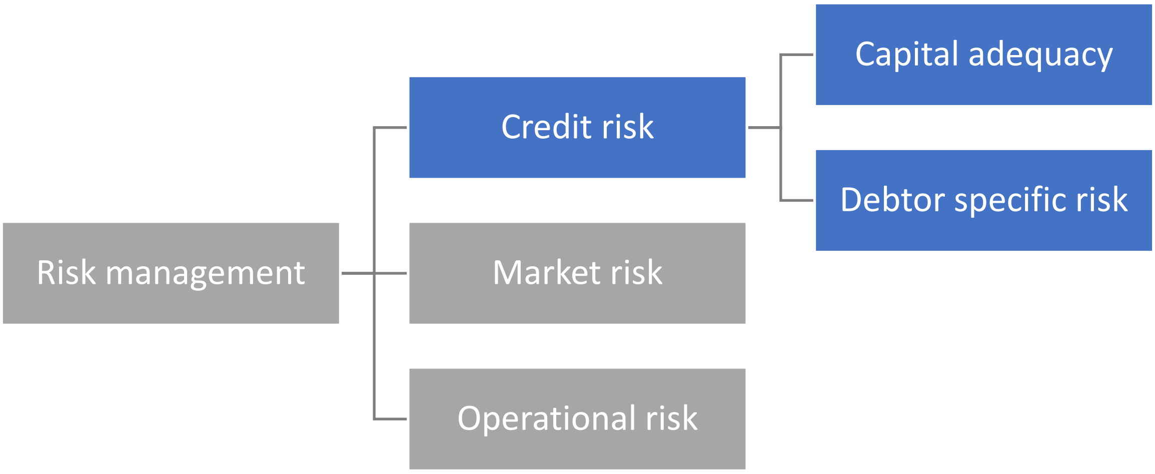


Figure 1. Theoretical framework and focus of the study marked with blue.

Methods used in forecasting LGD are linear regression, beta regression, random forest, generalized linear regression, and support vector regression. Prediction results from these models are estimated with different performance metrics and a ranking between the models is formed. Banking regulation will be evaluated from relevant parts in terms of internal LGD models. In LGD framework another concept is in-default model and expected loss best estimate (ELBE) which are required from IRB institutions. This thesis will not assess the in-default side of LGD estimation and is only interested in the LGD of performing loans. Also, the requirements relating bank's adoption of IRB approach are not discussed in this thesis.

1.4. *Structure of the Study*

Structure of this thesis is as follows. Section 2 discusses credit risk from a wider perspective in banking. Section 3 is focused on banking regulation maintaining focus on Basel framework and EU regulation. Section 4 assesses loss given default and past research on the topic. Methodologies and data used in this thesis are explained in section 5. Section 6 introduces the models employed in empirical LGD prediction and the related results. Section 7 concludes with suggestions for future research

2. Credit risk in banking

Bank’s risk assessment is critical for the institution itself and financial markets. As a disruption in a one bank sends negative shocks to others and the negative effect multiplies. (Koulafetis 2017, 10-13) Main risks of banks are operational, market, liquidity, and credit risks of which exposure to credit risk is the most significant. Credit risk is a relative concept which makes it difficult to standardize. (Win 2018, 744) It arises naturally from lending which has always been the primary business of banks. Successful lending requires banks to assess debtor’s creditworthiness with consistent accuracy. (Fight 2004, 1)

Bluhm et al. (2016) approach credit risk as actions for expected loss and unexpected loss. When given large pool of credit, losses are expected to arise. To cover these credit losses bank collects corresponding risk premiums from debtors in order to build capital buffer. This action has similarities to insurance business where costs of bad customers are covered with premiums collected from all customers. Bluhm et al. (2016) define unexpected losses as standard deviation from estimates e.g., expected losses. It is worth noticing that the IRB approach to credit risk is focused on unexpected losses from credits. The calculated risk weight output capital requirements for unexpected loss and the expected loss calculation are made under different practices. (BCBS 2021, 237)

2.1. *Definition of credit risk*

Saunders (2014) defines credit risk as “the possibility that promised cash flows on financial claims held by financial institutions, such as loans or bonds, will not be paid in full”. In other words, credit risk is the risk of borrower defaulting a loan or not fulfilling their financial obligations. Separating credit risk from other risks can be difficult as they may overlap. Unexpected change in market value of company owned assets affects the probability of default, in other words market risk contributes to credit risk. (Jarrow & Turnbull 2000, 272) Thus it can be argued that credit risk and market risk are not separable which can be the case for some assets. This kind of situation can happen in corporate bonds whereas retail credit risk is not as exposed to this interrelationship.

One key concept in credit risk quantification is the probability of default. It can be estimated from historical default rates or by other means. Other key parameter to quantification is recovery rate. Recovery rate is calculated as one minus loss given default. PD and RR are usually negatively correlated. (Koulafetis 2017, 210-233) Basel framework (BCBS 2021) also introduces exposure at default and effective maturity to be used in credit risk quantification. Exposure at the default is defined as the outstanding balance of debt at the moment of default. This EAD component is third major part of the IRB approach, but this thesis focuses on LGD.

Under the topic of credit risk counterparty credit risk is a risk that arises from the possibility of transaction counterparty's default before final transaction settlement. It is principally different from unilateral risk which lender is exposed as the lender only takes risk of losses. In counterpart credit risk both parties can have a positive or negative outcome. (BCBS 2021, 463) Counterparty credit risk arises usually in derivatives and other over-the-counter (OTC) trades. (Koulafetis 2017, 8-9) Due to these aspects this type of risk is not absent in retail credit institution. Counterparty credit risk will therefore be beyond the scope of this thesis.

2.2. *Credit risk management*

Managing credit risk is a multidimensional task and requires institution wide deployment. Major functions in credit risk management include shaping organization structure for complete risk assessment, modeling work, and monitoring. (Witzany 2017, 1-3) Starting point for a reliable risk framework is to identify and understand risks that the business is facing. Banks should only accept transactions that fit their strategy and portfolio profile. These decisions are culminated in the credit risk policy. It should document ways to measure, monitor, and control credit risk. (Koulafetis 2017, 13-14) It is sensible to assess these factors as they are direct representation of bank's strategy. From organizational perspective Witzany (2017, 6) argues that key idea in sound credit risk management is to separate business side from risk management. Doing so ensures that different remuneration can be implemented favoring good practices in both credit approval and setting exposure limits.

Credit risk mitigation is an important part of managing credit risk. Banks can mitigate credit risk with for example covenants, collateralization, and credit derivatives (Koulafetis 2017, 187). Collateralization is the main tool within scope of this thesis as retail credit has typically some type of collateral. It is a powerful way to mitigate risk especially with an easily liquidated collateral. If obligor defaults more liquid collateral typically means better recovery rates. (Koulafetis 2017, 193-194) The Basel framework sets out also credit risk mitigation ways for both standardized approach and IRB. Framework points out that mitigating credit risk can increase other risks such as legal, operational, or market risk. (BCBS 2021, 205)

2.2.1. Credit risk models

Corporate credit risk models can be divided into three categories: structural models, reduced-form models, and value-at-risk models (VAR). (Altman 2004; Scandizzo 2016) Structural model framework was developed by Merton (1974) and is based on Black and Scholes (1973) option pricing theory. Structural models assess company credit risk as a function of company's asset value volatility, interest rate and payout policy of both the firm and bond (Merton 1974). As company's asset value (market valuation) can be estimated and liability value is derived from the model, a default occurs when company's asset value is lower than the value of liabilities. According to Altman (2004), pay-off for bondholder in the case of default is the face value of bond minus a put option on the value of firm with strike price equal to face value of bond. Merton's (1974) structural model presents the concept of probability of default, although with different name.

The Merton model has been criticized for its strict assumptions which have been proven unfulfilled in empirical findings. Longstaff & Schwartz (1995) point out that assumptions for company to default only in maturity of debt and constant interest rates are unrealistic. This reasoning is easy to understand and shows flaws in structural-form models. Eom et al. (2004) tested Longstaff & Schwartz's (1995) model among other structural form models and found out that they did not produce high enough credit spreads for corporate bonds. To overcome these shortfalls reduced-form models were developed.

In contrary to structural-form, reduced-form models do not bind company default to value of its assets. Parameters related to asset valuation have to be estimated since they are non-observable. Reduced-form models see PD and recovery rate as independent from each other and exogenous variables. (Altman 2004) Reduced-form models turn the search of risk drivers from inside the company to outside factors (Zamore 2018). This thesis's model is built as reduced-form although predicted variable is not the traditional probability of default.

Credit value-at-risk models are aimed at measuring and minimizing potential losses by portfolio diversification. VAR models have produced many commercial solutions such as J.P. Morgan's *Credit Metrics* and McKinsey's *CreditPortfolioView* (Zamore 2018; Altman 2004). Among these observed model categories, most research and development has taken place for corporate credit risk. Consumer credit risk models have evolved to credit scoring methods. Credit scoring forms a scorecard to rank customer specific credit risk into a corresponding segment. Institutions determine a threshold value for scoring which is used to reject or allow the line of credit. (Thomas et al. 2005, 1007)

Credit scoring is well established method in consumer credit with utilizes in contacting possible customers, separating good loans from bad ones, and managing outstanding consumer debt (Crook et al. 2007). One common way to conduct this classification is logistic regression models. These models try to classify customers in good ones, who are allowed the credit, and bad ones who are rejected. (Thomas et al. 2005) Recent development has expanded credit scoring from binomial consumer selection task to debt characteristic determination. For example, scoring can be used to estimate maximum credit limit of credit card. (Crook et al. 2007)

3. Banking regulation

International banking regulation can be dated back to 1974 when Basel Committee on Banking Supervision was established. Basel Committee was founded by the Group of Ten (G10) countries to improve banking quality and to strengthen financial stability. Over the years it has published a framework of international banking regulation, known as Basel accord. (BIS 2021)

3.1. *Why regulation is needed*

Banking is one of the most regulated industry which is due to banks' critical position in society and financial markets. They act as financial intermediates deciding on who is allowed credit and at what price. (Koulafetis 2017,1) There are two main arguments for banking regulation although no consensus is formed on whether banks need regulation or not.

First pro regulation argument is the systemic risk to which banks are exposed. Santos (2001, 46) sees systemic risk as risk of bank runs. Bank runs happen when depositors simultaneously try to withdraw majority of their deposits. This can lead to problems as bank does not have all those deposits in liquid form and can thus lead to a bankruptcy. Santos (2001, 51) also points out that bank, as per usual all businesses are subject to adverse selection and moral hazard. These problems lead to need of information sharing and monitoring. Depositors have an incentive to monitor the bank but due to generally small holdings and scarce information monitoring is not done by individuals. Banking regulation can be used as coordinated monitoring to represent the depositors.

3.2. *Basel framework*

Basel Accord has set the main framework of banking regulation which is now adopted worldwide. Regulatory framework has been refined through history and the development is still ongoing. (BIS 2021) In Europe, compliance to Basel accord has been done by EU wide legislation of which European Banking Authority (EBA) implements. (EBA 2021)

3.2.1. Basel I

Basel I, published in 1988, made an important step towards international capital standards which where phased in by 1993. Its aim was to ensure that banks have enough capital to survive systemic shocks and to even out international competition. (Atkinson & Blundell-Wignall 2010, 10) Basel I targeted banks credit risk and required internally active banks to hold a minimum capital of 8% of risk-adjusted assets. (Santos 2001, 60) In the beginning, Accord was implemented only in G10 countries (BIS 2021).

The minimum capital requirement was further divided into Tier 1 and Tier 2 capital. Tier 1 consists of equity capital and disclosed reserves while Tier 2, also known as supplementary capital, consists of hybrid capital instruments, general loan-loss reserves, and subordinate debt. (Santos 2001, 60) Tier 1 capital is a good representation of core capital as it is mutual across jurisdictions, and it is wholly visible in financial publications (BIS 1988, 3).

Calculation of risk-adjusted capital was first introduced in Basel I. Banks assets were assigned into a one of four risk-classes. Risk classes were for example 0% for AAA rated sovereign debt and 100% for below B rated corporate debt. These risk-weights were then applied to adjust exposures according to their risk class. (Santos 2001, 60) Decided weights have a drastic effect on banks’ lending as 0% RWA does not require any additional capital. From capital management perspective regulation has a strong effect on lending policies. The minimum capital ratio (Tier 1 and 2) is calculated from the risk-adjusted assets (BIS 1988, 14)

Since its release Basel I has been adjusted and expanded to cover market risk. In 1996, Tier 3 capital was introduced to cover risk of losses from market movements. This addition offered banks a way to use internal models as an alternative to standard model in market risk determination. More specifically internal model was used to estimate value-at-risk. (Santos. 2001, 61)

Basel I has produced a wide range of scientific papers discussing its efficiency. Jones (2000) shows that banks have an opportunity to artificially lower effective risk-based capital requirements below the minimum of 8% set by the Accord. This can be achieved with securitization and financial innovations which alter the reported figures to look better than true situation is. Determined risk-weights were left intentionally broad to avoid micromanaging credit allocation. At this point regulation was quite simple to understand but it was too generalizing as only one risk-weight class saw collateral as risk mitigation factor. (Herring 2018, 185)

3.2.2. Basel II

The Basel Accord was designed to evolve over time which led to publish of a new refined framework Basel II in 2004 (BIS 2021). This framework consists of three foundation parts (referred as pillars): extended minimum capital requirements, supervisory review process, and market discipline (BCBS 2006, 6). In attempt to eliminate regulatory arbitrage opportunities, such as Jones (2000) point out, Basel Committee wanted to tie risk weights to banks actual exposures. A fundamental change from focusing on the borrower to evaluating risk based on third party credit rating agencies occurred. (Herring 2018, 189)

From the three pillars first one is the most important for this thesis. It consists of rules and methodologies on calculation of minimum capital requirements for credit, operational, and market risk. (BCBS 2006, 6) This capital has to be in place to buffer unexpected losses. Sum of risk-weighted assets considers all three risks which are estimated separately and pooled together with certain formula. (Atkinson & Blundell-Wignall 2010, 11) Supervisors started to see the original Basel Accord's simplicity as weakness. Thus, Basel II increased available risk weights and allowed banks to use their internal models also in credit risk estimation. Leading to added complexity of regulation and reducing transparency to the market. (Herring 2018, 190)

IRB approach is more complex method and can lead to more risk-sensitive risk-weights. In IRB approach bank is required to estimate PD, LGD, and EAD for each individual credit. (Atkinson & Blundell-Wignall 2010, 12) Comparison to standard model shows that calculation of risk-weighted capital is significantly more intensive and resource heavy under the IRB approach as risk-weights are not given externally.

Both the standardized and IRB approach have major flaws in them. IRB approach may be vulnerable to intended credit risk modeling with objective to avoid high capital requirements. On the other hand, standard approach relies on external credit rating agencies which can cause problems since some rating methodologies have shown under estimation of risks. (Herring 2018, 189) As risk-weights represent risk of individual credit not the portfolio they do not take into account portfolio diversification. Atkinson & Blundell-Wignall (2010, 12) point out that this behavior does not reward for portfolio diversification which by itself lowers risks. They also see the regulation framework to be pro-cyclical as credit policies tend to be loose in the good economic conditions and tighten in bad times.

3.2.3. Basel III

In the aftermath of financial crisis in 2007-2009 BCBS wanted to strengthen the Basel framework leading to publish of Basel III framework. It was a direct response to banks problems causing the financial crisis such as high leverage, small liquidity buffers, and poor governance. Basel III has been revisited many times after initial publish, most recently in 2019. (BIS 2021) New framework enhances banks risk coverage and strengthens minimum capital ratios by adding several new reserve buffers. For the first time Basel Committee set out standards concerning liquidity risks with liquidity coverage ratio and net stable funding ratio. (Nguyen 2019, 459) Former ratio promotes short-term resilience and the latter is more focused on longer term. (BCBS 2011, 9)

IRB approach to minimum capital requirements introduced in Basel II did not get any major updates. This was rather surprising as research has pointed out conceptual flaws in the IRB models. One major problem is the discrepancies between banks risk weights to obligors with similar characteristics. (Stupariu et al. 2019, 341-342) Current regulatory procedures can lead to different capital requirements for the same obligor in two different institutions. Stupariu et al. (2019, 346) argue that regulation for the validation of IRB models is faint. Historical validation data set should be long enough to have wide range of observations. Authors say that this data set is rarely based purely on internal data. To comply with the requirement institutions extrapolate internal data with external data. Using external data in the IRB methodology is against its founding ideology and mixing data from multiple sources can cause estimation error thus leading potentially to too low capital reserves.

3.3. *Capital requirements regulation in EU*

Capital adequacy and liquidity regulation framework in European Union includes Credit Requirements Directive (CRD) and Capital Requirements Regulation (CRR). (FIN-FSA 2020) This legislation is the direct implementation of Basel III framework in EU. Implementation started in July 2013 and since the start it has been reformed multiple times. Aim is naturally similar to Basel Accord, strengthening EU wide banking sector to be more resilient for economic shocks.

European banking authority develops more precise regulatory technical standards (RST) and guidelines on regulation implementation. (EBA 2021) For purpose of this study European legislation is more valuable since it is direct regulation for institutions in EU. Most important publishes are European parliament’s CRR (Regulation (EU) No 575/2013) and EBA’s (16/2017; 07/2016) guidelines for IRB approach.

3.4. *Regulatory capital calculation*

As discussed in previous chapters Basel framework sets rules and regulations on calculation of regulatory capital (BIS 2021). This thesis is focused on pillar 1 of Basel II which was later on reformed in Basel III. It covers capital, risk coverage, and leverage ratio of which risk coverage is most focused in this study. Banks have two distinct ways to calculate their regulatory capital: standard approach and IRB approach (BIS 2021). These methods approach the concept from two different aspects.

Total RWAs are calculated for market, operational, and credit risk (BCBS 2011). For this thesis market and operational risks are left out of the scope for reasons mentioned in chapter 1.3. For the following chapters different approaches are considered only from credit risk perspective.

3.4.1. Standard approach

Standard approach is the default model for RWA calculation. It uses standardized risk weights set by supervisors for different exposures. RWA is the product of corresponding risk-weight and exposure amount at given time. (BCBS 2021, 180) Standard approach is simplest of the three available methods since risk weights are given by supervisors. Obligors are divided into groups by their nature for example sovereign, corporate, and retail credits. These groups are further divided to sub-groups by assessing the credibility of corresponding obligor. (BCBS 2021, 187) Better creditworthiness equals to lower risk weight and lower regulatory capital reserve.

3.4.2. IRB approach

Banks and other financial institutions can apply for the use of internal models in calculation of RWA against credit risk. (European Parliament 2019/876, art. 107) Jurisdiction's supervisor, in Finland Financial Supervisory Authority, or European Central Bank can directly give permission to use IRB approach. Adoption of IRB approach is further divided into two separate philosophies. In foundation IRB banks estimate risk components internally but in some cases use external or given values to determine them. Generally speaking, in foundation IRB banks use their own PD estimate but rely on supervision estimates on other risk components. Second method is advanced IRB where banks must use their own estimates to all risk parameters. (BCBS 2021, 250)

IRB approach is divided in estimation of unexpected losses and expected losses. (BCBS 2021, 237) As these loss types differ significantly expected losses are left out of examination. Banks own estimation of unexpected losses is derived from risk components which are PD, LGD, EAD, and effective maturity. Basel Committee has defined standards on the components based on the underlying assets. These components are then used in different formulas for all assets to calculate the capital requirements. (BCBS 2021) For example retail exposures formula can be seen in appendix 1.

In the IRB approach banks categorize exposures to asset classes set by supervisor. Classes are set out to match underlying risk characteristics: equity, bank, sovereign, corporate, and retail exposures. Some classes are divided further into sub-classes, for example residential secured exposures and revolving retail exposures. (BCBS 2021, 238) These broad categories may be implemented in different ways as CRR (Regulation (EU) No 575/2013, art. 112) defines total of eight classes. This thesis will focus on the retail exposures. An exposure is qualified in retail class if debtor is a natural person or small and medium enterprise (SME). SME's size is not restricted in the legislation, but the total lending can be at maximum one million euros. (Regulation (EU) No 575/2013, art. 112)

3.5. *Future of banking regulation*

Basel Committee has released new reforms to Basel III framework in 2017. These reforms are sometimes called as Basel IV although BCBS sees them as finalizing reforms to the existing Basel framework. The new regulatory additions include revised standardized and IRB approach to credit risk, changes to market risk framework, and addition of interest rate risk to banking book. (BCBS 2017) Koch et al. (2017, 2-3) estimate that the imposed changes will have significant increasing impact on banks' capital. Authors estimate that in total EU banks will need a total of 120 billion euros more additional capital, assuming no mitigation actions by banks. The planned changes are phased in on multiple years between 2023 and 2028 (BCBS 2017, 12).

With the historical trend from Basel I to Basel IV it can be argued that banking regulation will keep on increasing. In the near future most important change for IRB banks will be the aggregated output floor. Output floor is the maximum advantage from IRB approach compared to the standard approach. (BCBS 2017, 11) In practice the output floor will restrict the risk-weighted assets calculated with IRB models. This is a major change as in previous regulation there has only been floors for individual estimates.

Output floor is calculated as percentage of RWA calculated with the standard approach. From 2028 onwards IRB RWA must be equal or more than 72.5 percent of the standard approach RWA. Change to output floor will be phased from 2023 onwards when floor will first be 50 percent and increase annually to the 72.5 percent. (BCBS 2017, 11-12) This change may decrease banks CET1 by 1.3 percentage when implemented and further putting pressure on return on equity (Koch et al. 2017, 2). The revised framework will also restrict the extend of advanced IRB models by disallowing its use for large corporations, banks, and other financial institutions. In conjunction with the output floor of total RWA, the components of RWs will see an increase in the minimum estimates. For example, retail mortgages will have a PD floor of 5 basis points and LGD floor of 5 percent. (BCBS 2017, 5-6)

4. Loss given default

Loss given default is one of the key components in minimum capital calculation for IRB institution. Since the addition of the IRB approach to Basel framework in 2004, banks have had the possibility to internally estimate LGD for their exposures. (BCBS 2006) After more than 15 years of active IRB banks, no market consensus has been formed for the topic. One reason for this is that the internal models are built specifically to the institution which makes the solution non-transferable. For the same reason IRB methodology is heavily regulated. Since supervisors allow banks to use their own tools and analytics they also need to ensure that the resulting models are robust (Schuermann 2004, 3).

In the IRB approach risk components have a linear development model (or life cycle) set out by the supervisor (EBA 16/2017, 11). Figure 2 showcases institutions internal processes of risk parameter estimates. This thesis will focus on the first part of the cycle: model development. For an approved IRB bank all steps are equally important as supervisors look at institutions as a whole and the IRB approach is a holistic model on risk assessment (Regulation (EU) No 575/2013, art. 143).

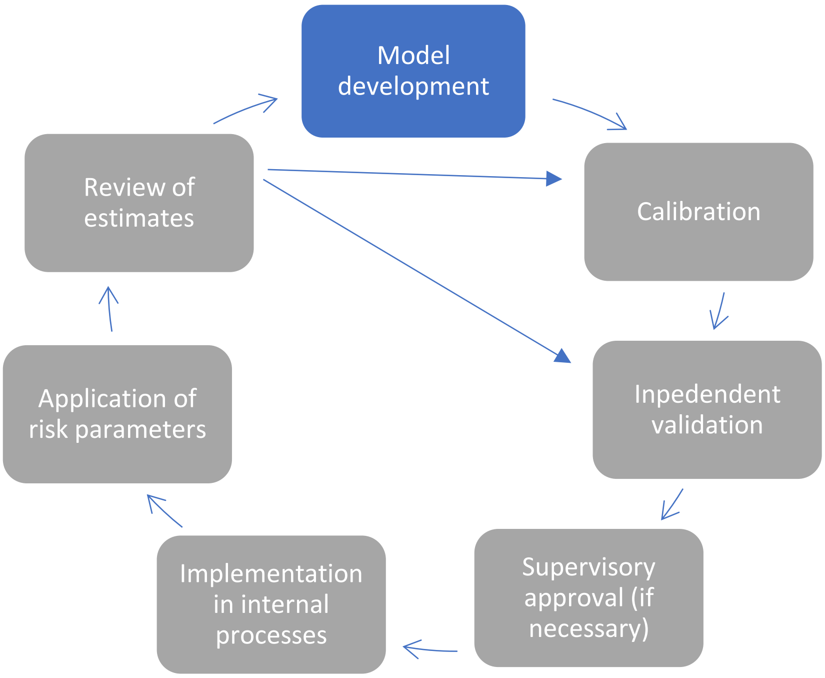

Model development
Calibration
Inpedendent validation
Supervisory approval (if necessary)
Implementation in internal processes
Application of risk parameters
Review of estimates

Figure 2. Internal cycle of risk parameter estimates. (EBA 16/2017, 11)

Past research on LGD has been tilted towards corporate loans and bonds while the retail credit has seen only a few studies. One reason for this is the availability of corporate data. For example, firms' annual financial reports are public whereas private individuals' information is not openly available. (Leow 2012) In addition the internal ratings-based method was not introduced until in 2004 in Basel II (BCBS 2006).

In context of bank capital regulation, loss given default is a multivariate problem, as LGD has to be estimated for performing (e.g., non-default) and in-default loans. For loans in-default, the best estimate of expected losses is also needed increasing total amount of loss estimates to three. (Regulation (EU) No 575/2013) Gürtler & Hibbeln (2013, 2358) differentiate LGD in-default and non-default models with the information gathered after default. The main factors being time in default and recovered cashflows since default. As per definition, performing loans do not have this information it is not considered in the proposed LGD model. ELBE models are also used to estimate losses for ongoing in-default loans and are beyond the scope of this theses. In spite of this research papers on these models are interesting as all these models are quite similar and EBA (16/2017, 35) emphasizes that transition of loan from performing status to defaulted should not result in inconsistences or discrete jumps in LGD estimates.

4.1. *Definition of loss given default*

Leymarie et al. (2018, 350) define LGD as potential loss occurring if the loan would default. In CRR, loss given default is defined as "ratio of the loss on an exposure due to the default of a counterparty to the amount outstanding at default (Regulation (EU) No 575/2013, art. 4 point 55). Definition of LGD is illustrated in Equation (1). In other words, LGD is the estimated loss ratio expressed as percentage of outstanding amount of debt at the moment of default. The amount of debt at default should not be confused to exposure at default (EAD) which is a different component in IRB minimum capital requirement formula. For convenience, in this thesis EAD is used to reference the risk parameter in the IRB risk-weight estimation. Term outstanding amount at default is used for the LGD context's obligation amount in the moment of default. Denominator of Equation (1) should include all capitalized interests, fees, and principal up to the moment of default (EBA 16/2017, mom. 131).

$$Loss\ given\ default = \frac{Amount\ of\ loss}{Outstanding\ amount\ at\ default} \tag{1}$$

Losses are defined in the CRR (Regulation (EU) No 575/2013, art. 5 point 2) as “economic loss, including material discount effects, and material direct and indirect cost associated with collecting on the instrument”. Economic loss represents the amounts of debt that obligor did not fulfill including principal, interest, and/or fees. Discount effects can impose losses as debt collection processes generate cash flow in multiple time points. (EBA 16/2017, chapter 6.3.1). Generally speaking, cash inflows to bank may come later than agreed if the obligor cannot pay in time, thereby making the highly uncertain.

Lowest amount of loss incurring from outstanding debt is zero meaning that the debtor fulfills their obligations and pays the principal, interest, and fees back in full amount. However, it is good to notice that in calculation of regulatory capital, Basel committee (2021, 282) has set an LGD floor to 10 percent for all loans. In the worst case, creditor can lose everything leading to loss of 100 percent. Hence LGD may vary between zero and one, both ends included. In some cases, it is possible for bank to lose even more than the amount of debt at default due to costs incurring from unsuccessful recovery proceedings (e.g. see Miller & Töws (2018, 192)).

4.1.1. Definition of default and recovery

Defining exact moment of default is crucial for LGD modeling. In this thesis default is derived from banking regulation. Capital Requirements Regulation (Regulation (EU) No 575/2013, art. 178) consider default to occur if either or both of the following has happened: an obligor is unlikely to fulfil their obligations in full or an obligor is due past 90 days of any financial obligation to the institution. The latter of these requirements is fairly easy to detect and monitor. The previous criterion is more subjective, and more information is needed to prove insolvency of the obligor. Corporate loans are naturally set to default if the company is in bankruptcy (Regulation (EU) No 575/2013, art. 178).

EBA (07/2016) has published regulatory technical standards on default definition which go into more sophisticated definition. These guidelines are not on the focus of this study as the earlier mentioned definition is accurate enough and covers most of the cases involved.

Schuermann (2004, 5) argues that all default scenarios do not end up in economic loss. For example, corporates might have cyclical sales with sales receivables longer than 90 days. In this case, such loans can be moved to default state but the payments to bank may take place in full later on. To comply with loans switching statuses too often EBA (07/2016, mom. 72) has set guideline to monitor defaulted loan for one year before it can be said to have recovered from default. In other words, if a sequence of default, recovery, and default occurs withing a year it should be handled as one default. Gürtler & Hibbeln (2013, 2355) say that recovery simply happens when default conditions are not met anymore. In that case an obligor has either paid back all arrears or a new payment plan has been settled.

After obligor defaults there are several types of losses that may be realized, such as loss of principal, workout costs from debt recollection process, and carrying cost of non-performing loans in terms of lost interest (Schuermann 2004, 6). Han & Jang (2013) use similar interpretation of losses. They specify that banks can utilize legal debt collection, for example foreclosures or injunctions. The process from default to partial or full write-off of loan can take up to several years.

4.2. *Estimation methods*

As basel framework or CRR does not specify a way to estimate LGD there are multiple statistical methods and model structures from which to choose. Most of the LGD models for performing exposures are work-out LGD models (EBA 2017). Work-out LGD means that the institution uses historical data from all available default exposures and identifies relevant risk-drivers (Leymarie at al. 2018, 350). Schuermann (2004, 6) points out that in work-out LGD, cash flows to distressed asset must be measured with care as they have a big loss mitigating impact. Thorough analysis of banks own defaulted credits leads to a model which can be used to forecast LGD for performing exposures. There are many studies on LGD modeling with different methods such as linear regression, tobit, beta regression, decision trees, and neural networks (Yaskir 2013; Leymarie et al. 2018).

LGD estimation models can be divided to linear and non-linear approaches. Both include a broad category of methods yet there are mixed results on the interpretability of the models. Linear regression is the most used model for LGD estimation (Miller & Töws 2018, 190). Zhang & Thomas (2012) used linear regression to estimate recovery rates for unsecured consumer loans in UK with good statistical results. Realized LGD values tend to have a bimodal distribution as values are concentrated to near zero and one (Han & Jang 2013; Zhang & Thomas 2012). This distribution suggests that more sophisticated method would be more accurate.

Loterman et al. (2012) studied LGD estimation methods and conclude that non-linear models are superior to linear ones. They tested 24 different methods for personal, mortgage, and revolving loans which makes their findings interesting for this thesis. For the datasets, artificial neural network (ANN) and least squares support vector machines were outperforming others. As there is no consensus on the best methodology to estimate LGD, the same applies to factors driving the amount of loss. Han & Jang (2013, 21) conclude that factors which have been found significant include size of loan, creditworthiness, product type, and collateral. Among these factors, collateral is the only one which is generally accepted as one of the most important factors. The authors think that the inconsistency is caused by differences in bank's loan portfolios, debt collection, and sample periods.

In forecasting potential losses from default, studies usually emphasize the amount lost. However total or partial recovery is equally important to consider as they have direct effect on realized LGD. Gürtler & Hibbeln (2013) propose a two-stage model structure with a probability of recovery estimator. They incorporated recovery as additional binomial classification to limited liability or unlimited liability and achieved performance improvements. For defaulted exposures probability of recovery can be derived from time since default or with significant reference points e.g., realization of collateral (Han & Jang 2013). It is worth noting that costs are not limited to only economic losses from lost principal or interest as Basel Committee distinguished two additional cost types which must be considered. One is indirect cost from the debt recovery process while the other is time span between the default date and actual recovered cash flows reflected by relevant discount rate. (Leymarie et al. 2018, 350)

Evaluating the goodness of LGD estimates can be a complex task. Leymarie et al. (2018, 349) conclude that usually LGD model comparison consists of three steps. First the sample of defaulted credits is split into test and training data sets. Secondly, all models to be compared are estimated on the training set. Finally, models are tested with the test set and the comparison is done on traditional methods such as mean squared error (MSE) or mean absolute error (MAE). This aspect leaves out the interrelationship of risk components while LGD is estimated individually. Further on it may be difficult to interpret the error in LGD estimate to the resulting capital requirement.

4.2.1. Calculation of realized LGD

Calculation of the response variable called, realized LGD, is included in the supervisor’s regulatory standards. The steps listed in this sub-section are necessary in calculating realized LGD for all defaulted obligations in the reference data set (RDS). EBA (16/2017, chapter 6.3.1) provides detailed descriptions on calculation of economic loss and realized LGD. To the CRR (Regulation (EU) No 575/2013, art. 4 point 55) definition of LGD, EBA adds the statement to include all amounts of principal, interest, or fees. These interest and fees should be incorporated into the calculation of realized LGD up to the moment of default. Interpretation of this is that they increase the amount of debt outstanding at default and affect both numerator and denominator of LGD ratio shown in Equation (2). Any fees or interest capitalized after default should not increase the amount of debt or economic losses.

$$Realized\ LGD = \frac{Economic\ loss}{Outstanding\ amount\ at\ default} \qquad (2)$$

The numerator of Equation (2) is calculated as the difference between outstanding amount of credit obligation at the moment of default and any recoveries after the moment of default. Realized recoveries to the credit obligation must be discounted to the date of default. (EBA 16/2017, mom. 132) The calculation of realized LGD can be confusing as it is divided into two differently calculated parts, but these share the same terminology. In particular, definition periods differ in the different parts as the numerator should consider additional drawings after default, whereas the denominator should not reflect any information after default. The discount rate to be used is set by EBA to be 3-month Euribor rate plus five percentage points (EBA 16/2017, mom. 143).

The last aspect to be reflected in the realized LGD calculation are direct and indirect costs. These are included in the calculation of economic loss (Equation (2)) and do not affect the credit balance at default. (EBA 16/2017, chapter 6.3.1.4) These costs have an increasing effect on economic loss and therefore, they are discounted in conjunction with possible cash flows from recoveries. EBA guidelines (16/2017, mom. 145-146) define direct costs as unambiguously attributable to a single specific debt collection. Examples of these are legal costs, insurances to collateral, and outsourced collection services. Contrarily, indirect costs cannot be attributed to individual obligations and therefore they are calculated at institution level and a chosen proportion of the total amount is assigned to individual obligations. They include all costs occurring from banks recovery processes, appropriate percentage of all ongoing costs, and any costs stemming from inhouse or outsourced debt collection which are not included in the direct costs.

A typical outcome of default scenario is a recovery and EBA has described technical standards on how to manage these outcomes. Guidelines state that those obligations which return to non-default state cannot be handled simply as a no loss outcome. (EBA 16/2017, mom. 135) This straightforward way would not consider the uncertainty of recovery and costs incurred from actions taken by the institutions. By applying the same treatment to cured cases consistency is confirmed across all observed defaults. EBA (16/2017, mom. 135) imposes a method of artificial cash flow calculation which considers all debt repaid when a loan is not anymore considered to be in default. The interpretation of this procedure is that cured obligations are treated with the same methodology although real life cash flows differ from the calculated ones. Outcome of a loan returning to non-default status should realize some amount of economic loss through costs and discounting of late repayments.

Capital requirements regulation (Regulation (EU) No 575/2013, art. 181 point 2(b)) specifies that institutions can reflect future drawings by the obligor after default either in the conversion factor or in loss given default estimates. Irrespective of the chosen method future drawings have to be considered in the economic loss calculation as discussed earlier. EBA (16/2017, mom. 139-142) states that reflection choice affects only the outstanding balance in the denominator of LGD ratio (Equation (1)). If institution chooses to include additional drawings after default in conversion factor estimate, additional drawings should be added on the denominator. In the opposite case outstanding balance is kept in original state. In this thesis, additional drawings are not added to the denominator as the choice concerns full IRB modeling, which is out of scope of this thesis and the choice should not affect the LGD estimation process.

One noticeable technical requirement in realized LGD calculation is handling of consecutive defaults. If an obligor recovers from default and again is classified to default state within nine months of the moment of recovery it has to be treated as one constant default (EBA 16/2017 mom. 101). In these such cases realized LGD should be calculated with the same procedure using the first default moment as reference point. Also, no artificial cash flow should be added at the moment of recovery. (EBA 16/2017, mom. 135) This treatment ensures that fast paced status changes are in line with observed history. It would be problematic to handle consequent default and recovery scenarios as separate because in the real business case they are treated as one default. To conclude all steps and required information in realized LGD an example calculation is shown in section 5.1.1.

4.2.2. Variable transformation

In LGD modeling it is not common practice to interpret independent variable discretizing (also called binning or bucketing) or other variable transformation (for example Han & Jang 2013; Zhang & Thomas 2012). These tools are utilized more in credit scoring, which is another form of credit risk assessment. Typically, credit scorecards are formed using variable binning by calculating information values (IV) and weight of evidence (WOE). (Yap 2011, 13277) Credit scorecards aim to separate good obligors who have a low chance to default from customers with high probability of default. (Thomas et al. 2005, 1007), thereby making them useful for assessing PD of loans. Despite different target and binary classification problem, practices used to determine optimal bins can be useful in LGD modeling, shown by Huajian & Tkachenko (2012).

Weight of evidence is described as logarithm of the ratio of “good” cases over “bad” cases in the particular group (Yap et al. 2011, 13277). Equation (3) illustrates this function. Positive WOE value implies that good cases are dominant, and the reverse is true for negative WOE values. SAS Enterprise Miner can form a scorecard based on WOE values and it is commonly used to conduct the analysis (SAS 2013).

$$WOE = \ln \left(\frac{P_{Good}}{P_{Bad}}\right) \tag{3}$$

where

WOE = weigh-of-evidence,

ln = natural logarithm,

P_{Good} = distribution of good cases,

P_{Bad} = distribution of bad cases.

Another key metric in scorecard formation is information value (IV) that is used to assess variables predictive power in separating good cases from bad ones. (Yap et al. 2011, 13277) These same principles are translated to R environment by Xie (2021). Using these techniques for the estimation of continuous dependent variable are rare but SAS (2012) has provided a guide for such purposes. It is not necessarily the absolute best technique, but scorecards are easy to understand and interpret in practice.

Matuszyk et al. (2010) implemented WOE-transformation to direct LGD response variable with good results. In this case the WOE equation is altered so that distribution of good cases is replaced with ratio of datapoints above or below average LGD (Equation (4)). Total number of datapoints above or below average LGD over the whole data set is the denominator in the equation. Interpretation of transformed WOE-values is that negative value is given for bins which have lower than average LGD. Average LGD is derived from the whole data set. WOE-transformation has many preferred reasons to use as it makes missing and outlier observation handling straightforward. Missing observations are grouped to their own bin and outliers are treated in the binning process so that they do not cause biases to estimates. A downside for the transformation is loss of information as variables are grouped to bins.

$$WOE = \ln \left({}^{\frac{n_a}{n_b}} \Big/ {}_{\frac{N_a}{N_b}} \right) \tag{4}$$

where

N_a = Total number of datapoints with above average LGD in the data set,

N_b = Total number of datapoints with below average LGD in the data set,

n_a = Total number of datapoints with above average LGD in the bin,

n_b = Total number of datapoints with below average LGD in the bin.

4.3. ***Regulatory requirements on data***

Data requirements for LGD models in the IRB approach are broad. CRR (Regulation (EU) No 575/2013) and EBA guidelines (EBA 16/2017) set out requirements for the data scope, period, and risk drivers included in the reference data set. Supervisor has strong motives to verify the data quality and consistency as data is one of the most important part in building internal models. Achieving robust models which fit banks credit portfolio institutions must make sure that the data is representative and of good quality.

One major concept in ensuring robust LGD model is data representativeness. Data representativeness framework is important because reference data is collected from at least five years before the estimation date. In this historical time period institutions portfolio, lending policies or debt collection etc. might have changed. (EBA 16/2017, chapter 4.2.3) If any change has occurred and model was built on data which does not represent institutions current portfolio the model performance can be faulty. Most critical aspects to research are changes in segmentation of exposures, definition of default, distribution range of key risk-drivers, and lending standards (EBA 16/2017, mom. 23).

Supervisor has long list of specifications for the data set used to build the internal model. First, the observation period should include all relevant information in the reference data set. “All relevant information” is a wide statement and EBA clarifies that all internally available data is relevant. (EBA 16/2017, mom. 147) Banks generate a lot of data from multiple sources, implying that in practice all data will not be used in the estimation of LGD. Arguably all identified risk-drivers are relevant, and these are discussed later. Chosen historical period should include a sufficient number of closed recovery processes as these are used for the calculation of observed (also called realized) LGD values (EBA 16/2017, mom. 148). Incomplete recoveries are not used for the LGD model data set but those must be included in the LGD calibration exercise which consists of long-run average and economic downturn LGD values.

According to EBA (03/2016, art. 51) institutions must include all complete defaults in the RDS. Another key requirement which concerns the defaulted loans is to measure and include all withdraws and payments from the debtor between default and final resolution (EBA 16/2017, art. 108). In other words, all cashflows must be recorded before default and until the final settlement of loan. Recording cashflows can be difficult as some debt collecting processes can take up to several years. From modeling purposes, it may become problematic as loan obligations final expiry is typically after 15 or 20 years (Enforcement Code, 705/2007, chapter 2 §24-27). To comply on long debt collection processes EBA specifies that institutions should calculate a maximum length for recovery process for different types of exposures. It must be specified so that vast majority of recoveries are included, and extra-long cases e.g., outliers are left out. (EBA 16/2017, mom. 156) Guideline on maximum length is set out only for calculation of long-run average LGD which is used in LGD calibration. As it is calculated from observed realized LGDs, the guideline can also be followed for the estimation of LGD values.

Regulatory framework sets out risk-drivers which institutions need to consider. These are transactional-, obligor-, institution-, and external-related risks. (EBA 16/2017, mom. 121) It is not determined whether the analysis must be quantitative or qualitative but inclusion to RDS is one straightforward way to assess them. Transactional risk includes type of product, collateral, loan-to-value (LTV) and exposure size. The obligor-related risks are for example geographical region and industrial sector for businesses. (EBA 16/2017, mom. 121). These are easy to understand and have been used in past research by Yashir & Yashir (2013) and Thomas (2005). Internal risks which include governance and possible merges are more complex to evaluate and are left out of this study (EBA 16/2017, mom. 121). From external factors macroeconomic variables are interpreted in many LGD models and included in the LGD models examined in this thesis. All these potential risk-drivers have to be analyzed at the moment of default and at least a year before (EBA 16/2017, mom. 122). Institutions can therefore pick a reference date for these risk-drivers with a year before default based on their view of best representativeness.

One of the main risk-drivers for LGD is amount and type of collateral. (Han & Jang 2013, 21) and therefore, regulation has strict rules and practices on collateral management in terms of LGD estimation. First of all, collateral type can be implemented in the model directly as risk-driver or as a segmentation parameter. Internal management, valuation and policies can differ with different types of collateral, and these should be analyzed and documented. (EBA 16/2017, mom. 126) Also, debt recovery process characteristics vary over different types of collateral. For example, real estate realization is completely different from selling securities. By analyzing historical recovery patterns of different collaterals institutions can group homogenous types into same groups. Institution’s attention should be on biases arising from collateral which covers multiple exposures and exposures secured only by part of the collateral. To counter these biases institutions must have a proved allocation method for parts of collateral to corresponding exposures. Also, data set should include information of the full collateral value even though only part of it is used to secure exposures. (EBA 16/2017, mom. 128-129)

5. Research methods and data

In this section data and methodologies of this thesis are covered. Firstly, data sources, processes and nature are addressed. Secondly, all statistical methods used on LGD estimation are explained.

5.1. *Data*

The data set consist of retail credits based in Finland which have defaulted at some point in their lifetime. To be included in the data set loans must have defaulted and resolved within the period. All data preprocessing and modeling is done in R. Data is built in a way that each default case is unique, but a credit obligation can exist multiple times if there has been more than nine months between the separate defaults. Total amount of default observations is 3 202 and the sample period ranges from January 2006 up to March 2016. Cashflows after default are observed for maximum of five years to calculate the realized loss of obligation which is used as the response variable. Predictor variables are collected six months before default.

5.1.1. Response variable calculation

Realized LGD i.e., the response variable is calculated according to regulatory requirements discussed in section 4.2.1. An example of calculation is illustrated in Table 1. The example case is a 100 000 € debt backed up by collateral. Obligor has defaulted and proceeded to legal debt collection. The collateral was successfully realized withing the maximum observation period of five years. Outstanding balance at default is the only value in the denominator of LGD Equation (2) whereas it is a starting point in the economic loss calculation in the numerator. It includes all principal, interest, and fees up to the date of default, but it does not consider anything from that point onwards.

Table 1. Example case and steps on calculating realized LGD.

Type of cashflow	Amount	Discounted amount
Outstanding balance at default	100 000€	100 000€
Drawdowns (+)	1000€	950€
Direct costs (+)	500€	480€
Indirect costs (+)	300€	290€
Outstanding balance at default (numerator)	101 800€	101 730€
Recoveries (-)	60 000€	57 400€
Economic loss	41 800€	44 300€
Outstanding balance at default	100 000€	100 000€
Loss given default	42%	44%

The calculation of economic loss must include relevant cash flows also after default. They are reflected as additions or reductions to the debt balance. Additional drawdowns, direct and indirect costs are discounted to date of default and added to the outstanding balance at default (numerator). On the other hand, recoveries realized reduce the balance after discounting them to the same date. Remainder of these components is economic loss. Final realized LGD value is then calculated as ratio of economic loss and outstanding balance at default, with no reductions or additions. Special cases in the data set are debt collections which last longer than five years. Around 10 percent of observations are longer than five years. As in business sense long-lasting debt collection is unwanted these cases are treated with conservatism. In practice, the amount not recovered at that point is considered to be lost fully.

5.1.2. Summary of data and its nature

The data set represents retail and small enterprise loans with amounts presented in Table 2. Modeling period consists of ten years from January 2006 to March 2016. Around 60 percent of observations are from private customers and 40 percent are SME companies and entrepreneurs.

Table 2. Loan types and amounts in the data set.

Retail consumer credit	1 110
Consumer housing loan	820
Small enterprise loan	780
Flex credit	195
Company others	165
Consumer others	130
Total observations	3 200

Variables included in the data set are chosen to fulfil regulation requirements e.g., regulation (EU) No 575/2013. Variables are broadly categorized as obligor related, obligation related, and collateral related. Full list of variables can be found in Appendix 3. Examples of obligor related information are the total amounts of deposits, time as customer, and business industry. Credit balance, type of obligation, and number of debtors are variables related to the specific obligation. Data from collateral consists of sufficient collateral amount (value of collateral after hair cut), nature, and market value of collateral. Gathered exploratory variables are in line with previous research such as Matuszyk et al. (2010).

The observed historical LGD is distributed with peaks at zero and one which is typical for LGD (Han & Jang 2013; Zhang & Thomas 2012). The peak at zero is greatly higher than that at one (Figure 3). Mean and median for observed values are 12% and 0% respectively. This implies that most of the observed defaults are settled by obligor paying their late payments. In total, 37 percent of cases resulted in a loss, and only 12 percent in a loss of 50 percent or more. LGD values are very low implying that the underlying credit portfolio is of good quality and debt collection processes are well established.

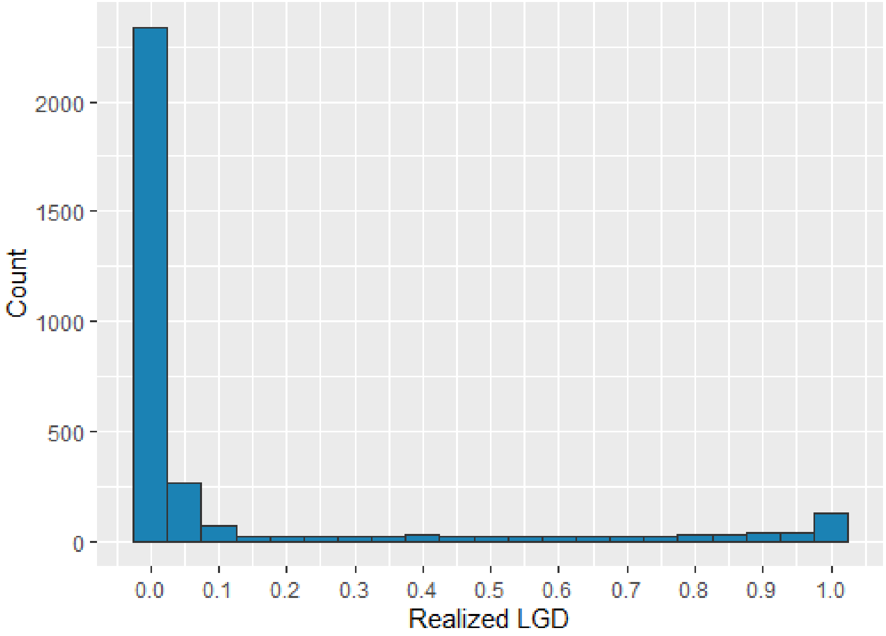


Figure 3. Histogram of realized loss given default.

Data set consists of long historical period with a growing trend of observations as can be seen from Figure 4. Year 2016 has fewer observations because it is not a full year as others are. Observed defaults are recorded on a monthly basis but as the yearly total is quite low the variance between months and years is large, implying that the outstanding amounts of loans, types of collateral, and obligors vary across the period. This variation will be taken into consideration in the out-of-sample performance tests and in the splitting of data into training and testing samples.

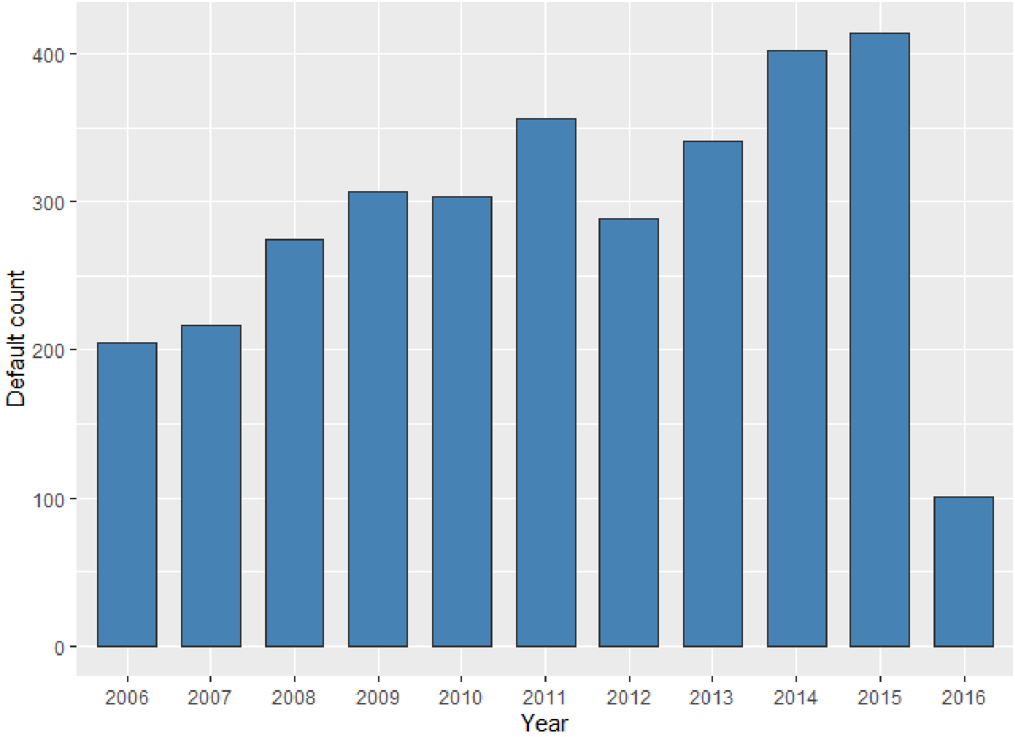


Figure 4. Observation counts per year in the data set.

Variable correlation was examined with WOE transformed variables. Full correlation matrix is shown Appendix 2. Transformed variables are used for the convenience of comparison as they are directly comparable. Former applies also if the original variables were categorical and continuous type as they are compared on the same scale. From the matrix it is clear that the variables are not strongly correlated. However, the variables that are related to collaterals are exceptional in this respect. Few highly correlated variables are not a problem as they all do not end up in the model to avoid multicollinearity. Realized LGD is weakly correlated to independent variables. Strongest positive correlation is to total haircut value of collateral.

Visual analysis reveals that some independent variables have a good predictive power. Figure 5 shows a boxplot of realized LGD and number of debtors. It is clear that obligations with one or two debtors have a wide range of realized LGD values although their median and mean are very close to zero. Dots in the figure represent observations that are more than 1.5 times the interquartile range (IQR) away from median. In other words, they are classified as outliers. As discussed, earlier LGD data sets are very skewed which cause difficulties to the estimation process. In Figure 5, cases with more than two debtors have significantly narrower whiskers. This is due to the lack of observations with these amounts as three and four debtors have seven and six observations, respectively. The variable shows discriminatory power as realized LGD values decrease when the number of debtors increase. Observations with one debtor have mean realized LGD of 0.13 while it is almost halved to 0.08 when there are two debtors.

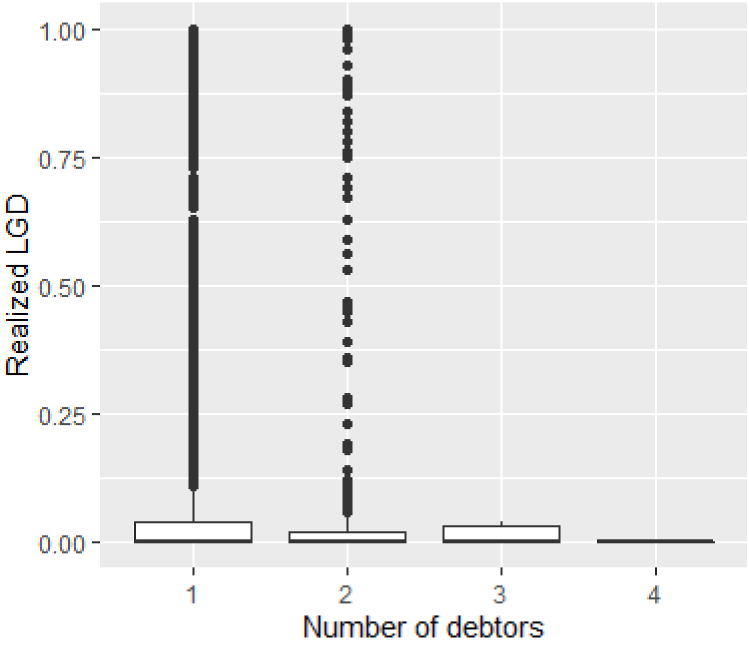


Figure 5. Boxplot of number of debtors and realized LGD.

Continuous numerical variables show lesser correlation to the realised LGD. For example, scatter plot of issued amount and realized LGD is illustrated in Figure 6. Dots are little transparent to showcase better the clustering density. Blue line in the figure represents the regression line between the variables. It is with a small upward slope meaning that larger issued amount leads to slightly bigger LGD values. This trend is very hard to see without the line e.g., it is very weak. The relationship comes clearer when the issued amount is discretized to bins. This process is discussed in detail in section 6.1.

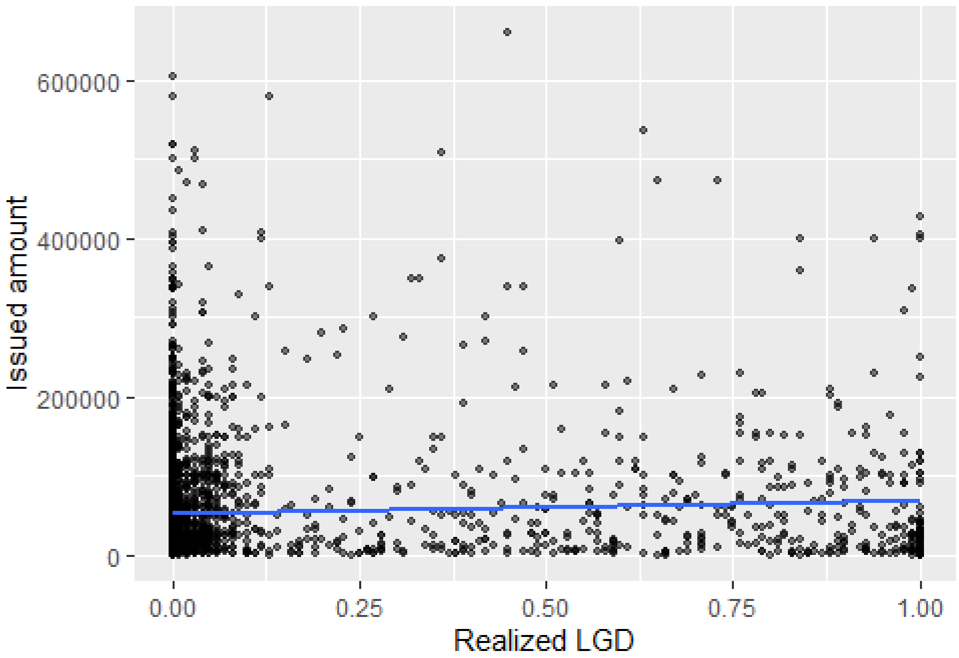


Figure 6. Scatter plot of issued amount and realized LGD.

5.2. ***Methodologies***

All methods used in this thesis are statistical ways to analyze whether a variable has causal relationship between one or more variables (Yan & Su 2009, 2). Past research covers a wide range of methodologies to forecast LGD. This study uses five different methodologies which are explained in the next sections. These methodologies were chosen because of their favorable applications in former research papers, and their implementability in reasonable time and effort.

Multiple ways of forecasting LGD have been employed in earlier literature, including direct estimation, recovery rate estimation, and two-step combinations of both. The purpose of this thesis is to directly estimate LGD. The output from all the models is directly the LGD level of which accuracy and calibration is evaluated.

5.2.1. Linear regression

Linear regression and its variants are widely used and studied methodologies which are common in many fields of studies. Multiple linear regression is typically equated in the form shown in Equation (5). It is a method to fit a linear line in the data which represents its characteristics best. (Yan & Su 2009, 11-12) Fitting of the line can be done in a few different ways of which least squares method is used in this thesis. Least squares method evaluates squared distances from observed data points to predicted values. The best possible model finds regression coefficients that minimize the sum of squared distances. (Yan & Su 2009, 11-12) This evaluation is called mean squared error and it is an established method also in LGD forecasting (Leymarie et al. 2018, 349).

$$y = \beta_0 + \beta_1 x + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon \tag{5}$$

where

y = response variable,

β_0 = intercept term,

β_k = regression coefficients,

x_k = predictor variables,

ε = error term.

In the linear regression the response variable is by definition not restricted. In LGD context this is problematic since LGD is limited to have values between zero and one, both ends included. One simple solution is to estimate LGD values with regression and take positive estimates as such. Negative values can be assigned to zero to fulfill the definition. (Yaskir & Yaskir 2013, 27-30) This method is very easy to interpret and does not cause problems in the model estimation. If estimated values are deeply negative then the model can be flawed. Yaskir & Yaskir (2013, 27-30) use censored regressions such as the Tobit model. The Tobit model suits LGD estimation very well as it is used to model dependent variables which are censored in some way. LGD can be seen as left-censored as negative values are not applicable. Situations resulting in loss greater than the outstanding balance are practically comprehensible.

5.2.2. Generalized linear regression

Generalized linear regression is an extension to the ordinary linear regression. GLM models are used to predict dependent variables which are not normally distributed. This is done via a related link function. (Olive 2017, 389-393) GLM with a gamma distribution (sometimes referred as gamma regression) is applicable to dependent variables which are gamma distributed. The LGD data used in this study is visually analyzed to be close to gamma distribution. Dunn & Smyth (2018, 445-446) propose that gamma GLM is suitable for positive continuous data with right skewness. For this reason, gamma GLM is tested in the estimation process.

Commonly used notation of probability function of gamma distribution is shown in Equation (6). The shape α and scale β parameters are defined to be greater than zero along with the dependent variable y. Shape, scale parameters and variance term, which is constant in gamma distribution, shape the nature of the distribution. (Dunn & Smyth 2018, 427-431) To comply with the restricted values of realized LGD, observations with zero realized LGD can be adjusted with a small constant to be above zero.

$$P(y;\alpha,\beta) = \frac{y^{\alpha-1}\exp\left(-\frac{y}{\beta}\right)}{\Gamma(\alpha)\beta^{\alpha}} \tag{6}$$

where

α = shape parameter,

β = scale parameter,

Γ = gamma function.

Gamma GLM has several link functions which can be used. A logarithmic link function is most common while other link functions available are called identity and inverse functions. Most suitable link function is challenging to know beforehand implying that testing different functions is a sensible strategy to find the best one. (Dunn & Smyth 2018, 433-436) All usable link functions will be tested in the thesis.

5.2.3. Beta regression

Beta regression is a regression technique proposed by Ferrari & Cribari-Neto in 2004. It is used to model a variable in the open interval of zero to one such as rates, proportions, and concentration. Thus, it is commonly used in LGD modeling. Beta regression uses parameterization of beta density with a precision parameter and variate mean (Cribari-Neto & Zeileis 2010). A common function for beta distribution is shown in Equation (7). Parameters p and q are defined to be positive values.

$$f(y; p, q) = \frac{\Gamma(\mathrm{p+q})}{\Gamma(\mathrm{p})\Gamma(\mathrm{q})} y^{p-1}(1-y)^{q-1} \tag{7}$$

where

p, q = shape parameters,

Γ = gamma function.

Similar to generalized linear regression beta regression uses link functions. These include logit, probit, and logarithmic links, for example (Cribari-Neto & Zeileis 2010, 3-4). Different link functions give flexibility in the modeling process as their relative performance can be tested rather easily. Beta regression does not typically use raw residuals due to heteroscedasticity of the model. Instead, Pearson’s residuals are used as Ferrari & Cribari-Neto (2004) propose.

5.2.4. Random forest

Decision trees are models which process data to split it into groups with high probability of event outcome (Olson & Wu 2020, 57). Tree-based models are quite common and variations occur. One evolved method is random forest (RF). It is a nonparametric method of statistical learning introduced by Breiman (2001). RF and other tree models are suitable for classification and regression problem settings (Olson & Wu 2020, 57). Estimating LGD is a regression problem and RF is used in previous research (Miller & Töws 2018). Tree models are constructed in the form of set rules which are represented as tree nodes. Outcome of the tree is represented as end nodes, called leaves. (Olson & Wu 2020, 57)

A typical decision tree model grows one tree e.g., the model. Idea behind RF is to aggregate results of hundreds or thousands of decision trees. (Genuer & Poggi 2020, 33-34) Use of RF instead of decision tree in this study is based on the added stability to the resulting model. The data set is rather small and a slight chance in it could harm the predictive power of decision tree. As random forest algorithm grows a diverse set of trees the resulting model is much more stable. More trees in the RF model training and randomness in their formation improve prediction performance. Most common aggregating function in RF is called “random inputs”. In this method a pre-set number of variables is randomly selected to build the tree. From this sample the best way to split it is identified and results from these split nodes are aggregated. (Genuer & Poggi 2020, 34-40)

Machine learning models typically need tuning. In the case of RF tuning is rather simple as there are only two meaningful parameters to tune. First, the number of trees is defined i.e., how many iterations are used to aggregate the model. The rule of thumb is the more the better. (Genuer & Poggi 2020, 43-44) Higher number of trees is used to reduce the prediction error but as number of trees grow the benefit diminishes. The second parameter is the number of variables in a node which is much more important. Different number of variables used can change the prediction significantly. A commonly-used starting point for regression trees is one third of total number of variables in the data set. (Genuer & Poggi 2020, 44-46)

5.2.5. Support vector regression

Support vector regression is a generalization of support vector machine (SVM) classification problem. SVR is designed to handle regression problems, thus it is used in LGD research. An SVM optimizes classification by finding a maximum margin around hyperplane and simultaneously correctly classifying observations. The margin is measured with support vectors, hence the name. Generalization to SVR is achieved with a ε-insensitive region called the ε-tube. (Awad & Khanna 2015, 67) SVR can be used to both linear and nonlinear problems.

SVRs build a regression model based on an optimization problem Equation (8). In modeling nonlinear relations, the optimization is mapped to higher dimensional space known as kernel space. In Equation (8) φ represents the feature transformation to kernel space. (Awad & Khanna 2015, 72-73) This approach will be used later on the study as LGD's relation has been researched with nonlinear models with promising results (Loterman et al. 2012). Possible loss functions in the SVR optimization are linear or quadratic. (Awad & Khanna 2015, 69) As with the other model's loss functions these will be tested to find the best fitting one.

$$\min \frac{1}{2} \left| |w| \right|^2 + C \sum_{i=1}^{N} \xi_i + \xi_i^* \tag{8}$$

subject to

$y_i - w^t \varphi(x_i) \leq \varepsilon + \xi_i^* \qquad i = 1, \dots, N$

$w^t \varphi(x_i) - y_i \leq \varepsilon + \xi_i^* \qquad i = 1, \dots, N$

$\xi_i,\ \xi_i^* \geq 0 \qquad i = 1, \dots, N$

where

w = vector of independent variable values,

ε = threshold value,

ξ_i, ξ_i^* = slack variables.

The minimum of Equation (8) is calculated with partial derivatives and by using Lagrange multipliers α^* and α (Awad & Khanna 2015, 70-73). The final regression formula takes the form shown in Equation (9). Model specific kernel function is represented with k. Bayesian linear regression is closely linked to SVR as it builds regression models in similar way using Bayesian inference (Awad & Khanna 2015, 74). These types of regression models are not used in this thesis due to the similarity of methods.

$$f(x) = \sum_{i=1}^{N} (\alpha_i^* - \alpha_i) k(x_i, x) \tag{9}$$

where

α^*, α = Lagrange multipliers,

k = kernel function.

5.3. ***Performance evaluation***

Model's performance is evaluated with qualitative metrics from perspectives of estimation accuracy and discriminatory power. Loterman et al. (2012) used this evaluation scale, and it fits the LGD context perfectly. Alongside quantitative methods, qualitative measures are used to evaluate results visually.

The employed performance metrics used are mean absolute error (MAE), mean squared error (MSE), coefficient of determination (R^2), Pearson's correlation coefficient, and area under ROC curve (AUC). MAE is calculated as total sum of absolute error between the predicted LGD and realized LGD. MSE statistic is similar but measures prediction errors as squared distances. For both of these metrics lower value is considered better as then the prediction error is smaller. MSE may be more generally used but MAE is very usable in LGD error measurement as the target variable is bounded between zero and one, both ends included. Interpretation of MAE is also more convenient as the metric is on the same scale as the target variable.

R^2 measures models accountability to the variance of target variable (Draper & Smith 1998, 220-221). Maximum and the best value of R^2 is one and the worst is zero. Interpretation of the metric is the proportion of variance that the model can explain. R^2

is calculated as one minus ratio of residual sum of squares and total sum of squares (Draper & Smith 1998, 220-221). AUC is typically used to evaluate classification models meaning it is not directly applicable to LGD estimation. Loterman et al. (2012) used the metric to evaluate model’s performance on separating observations below the average from those higher than average. This discriminatory power is especially good in the data set used in this thesis as the average LGD is 12 percent which is really close to the LGD floor set by supervisors (BCBS 2021, 282). Finally, correlation of predicted values and the target variable is evaluated.

6. Estimating loss given default

This section will go through the LGD modeling process. First data set transformation is explained, then all used models are discussed in detail. Lastly, modelled results are evaluated, and performance ranking is made.

6.1. *Variable transformation*

Transformation was carried out to all independent variables as discussed in section 4.2.2 with the formula shown in Equation (4). Determination of number of bins and corresponding cut points is a process which aims to create few discrete bins with differing risk characteristics. Once the bins are formed and optimized, they are applied to the data set so that all original values are transformed to the same WOE scale. Observations are assigned to the WOE values based on the corresponding bin cut points.

Process of variable binning starts from inspecting corresponding variable distribution in the whole data set. Let us use time as customer variable as an example case. Time is measured in years so that the minimum value is one and the oldest observation has been client for 47 years. Mean and median for the variable are 12.36 and 10.00, respectively. To see relationship between time as customer and realized LGD the variable is cut to ten deciles and average observed LGD is calculated for all deciles. Figure 7 shows that there is a downward trend on LGD when time as customer gets longer with couple of expectations. For the final binning, its preferred to be monotonic growth or decrease in riskiness. In dealing with deciles [7,9), [11,14) and [16,20) one solution is to reduce the number of bins. Pearson’s correlation coefficient is -0.07 for the variable with the realized LGD.

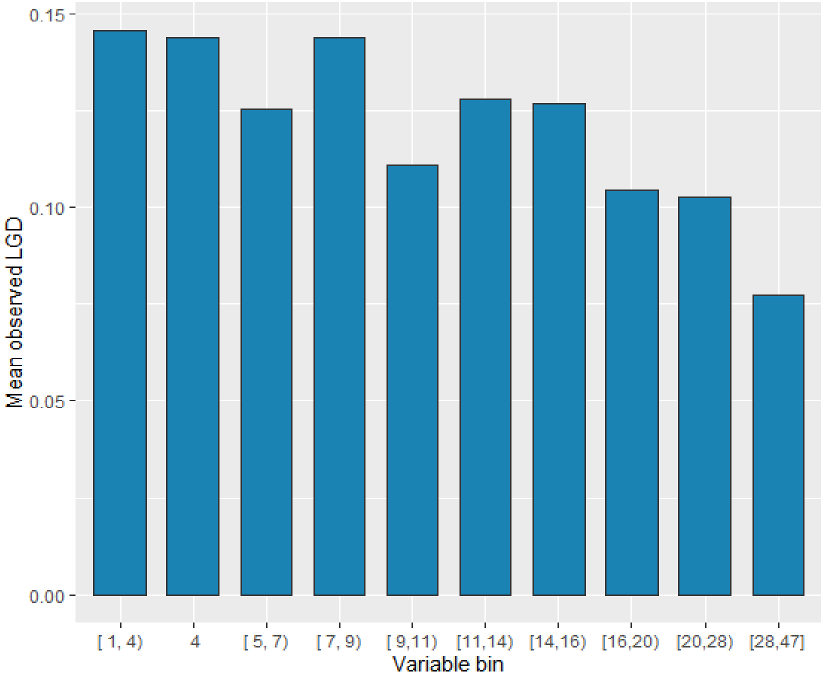


Figure 7. Ten deciles of time as customer and their corresponding average observed LGD values.

Finding of optimal bins is an iterative process. There are algorithms designed for it but for this thesis the binning process is done manually. The achieved final binning for time as customer is shown in Figure 8. It can be seen from the binning that LGD is now clearly monotonically decreasing as age increases. If the variable would have missing values they would be assigned to special bin for all missing observations. This was the case for couple of other variables. The next step in variable transformation is to calculate WOE-values for all five bins. Calculation is carried out as Equation (4) shows. Average observed LGD for the data set is 12.19 percent and in total 542 of observations are above the mean and 2 667 lie below the 12.19 percent.

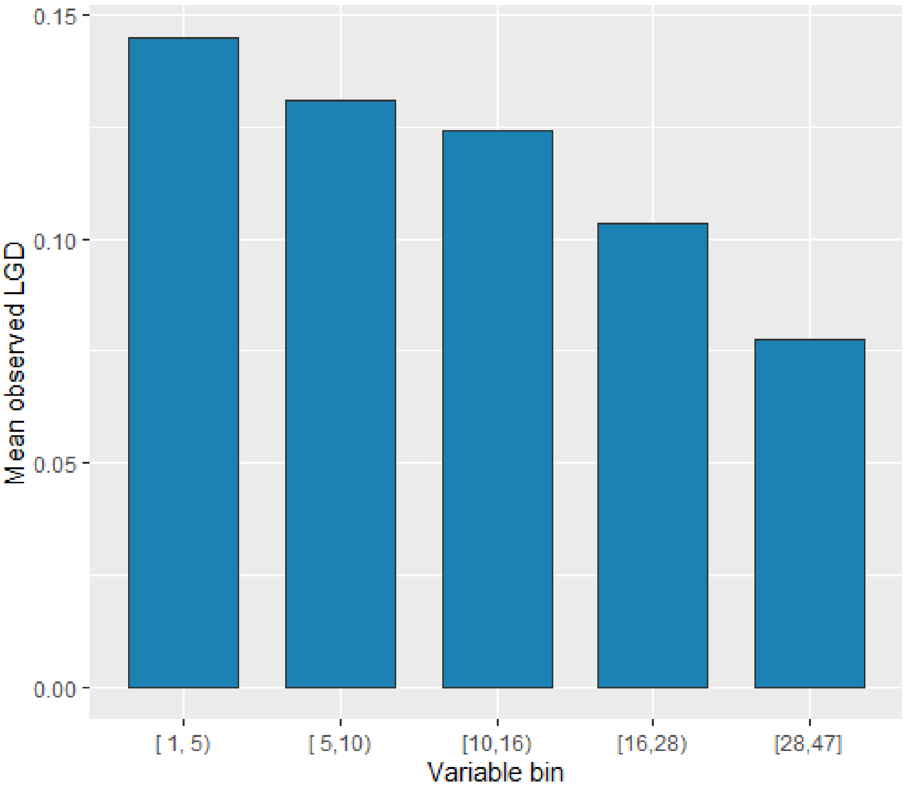


Figure 8. Final binning of time as customer variable and bin's average observed LGD.

Complete information of transformed time as customer variable are shown in Table 3. Mean LGD levels vary between 14.5 and 7.7 percent showing good discriminatory power of the variable. The WOE value column shows corresponding WOE value which is applied to all observations between the lower and upper limit of the bin. The last column of Table 3 shows that all bins have healthy amount of observations in them. To conclude, time as customer is a good example of transformed variable as some variables are more difficult to split into bins.

Table 3. WOE bins and statistics of time as customer variable.

Bin (years)	Mean LGD	WOE value	Count of observations
[1, 5)	14.49%	0.23	685 (21.4%)
[5 ,10)	13.09%	0.13	850 (26.5%)
[10, 16)	12.44%	-0.02	717 (22.3%)
[16, 28)	10.35%	-0.23	662 (20.63)
[28, 48]	7.74%	-0.51	295 (9.19%)

6.2. ***Model structures***

The models built to estimate loss given default are discussed in this section. The models are estimated with the transformed variables. From the initial data set, no variable reduction has been done. The individual models' variables are included or excluded based on their respective significance and model performance. All models are trained on the data set same split of 70% for training and 30% testing sample. All metrics are based on the testing sample.

6.2.1. Linear regression

Linear regression model is built on iterative variable selection. Process begins from fitting a linear model with all the independent variables. Next, a stepwise regression is applied with R's step-function. Stepwise regression forms subsets from the starting variable group by removing or adding one variable from the selection. New subset of variables is tested, and the algorithm continues until the best model is selected based on Akaike Information Criteria (AIC). Stepwise regression resulted in the variables shown in the fourth column of Table 4 plus number of debtors. However, the number of debtors had a p value of 0.07 meaning it is not statistically significant at 95 percent confidence level and was left out of the final model. To confirm the variable selection of stepwise algorithm various variable combinations were tested. Testing underlined the selection as no other combination performed better.

Table 4. Initial linear regression and the final model variable names, coefficients, and p values.

	Initial model		Final model	
Variable	Coefficients	P value	Coefficients	P value
Intercept	0.142	0.0000	0.142	0.0000
Amount of deposits	0.021	0.4843		
Business industry	0.095	0.0002	0.099	0.0001
Collateral type	0.070	0.0000	0.085	0.0000
Customer segment	0.019	0.3391		
Haircut value of collateral bundle	0.068	0.0000	0.082	0.0000
Issued amount	0.052	0.0001	0.055	0.0000
Number of debtors	0.035	0.0827		
Number of guarantees	0.017	0.5305		
Number of pledges	0.013	0.6571		
Percentage of paybacks	0.065	0.0003	0.061	0.0006
Time as customer	-0.011	0.6737		

Prediction results from the linear regression are promising. R^2 of the model is 0.132 which is on the lower end but for LGD models R^2 value is typically low (Loterman et al. 2012). MSE and MAE are 0.071 and 0.173, respectively. It seems that the WOE transformation has transformed dependent and independent variables relationship to linear form which suits linear regression. Predictions are problematic in the sense of prediction range as minimum forecasted value is -0.04 and maximum 0.48 compared to actual values which range from zero to one. Small negative values are not a problem as they can be rounded to zero to imply no loss for the obligation. The model cannot forecast highest LGD values, but the mean of estimates is 0.12 which is the same as observed values. Similar situation is replicated with the other models, and it may stem from the highly skewed data as only around 10 percent of observations have a realized LGD of 50% or higher.

Linear regression shows a satisfactory discriminatory power with an AUC value of 0.730. Values between 0.7 and 0.8 can be considered as acceptable. Correlation coefficient between predicted and observed LGD is 0.36. Overall, linear regression shows promising results in LGD modeling as was expectable. These results can be thought to be the baseline for other models as linear regression is a popular method in earlier LGD research. Due to its popularity among other fields, it is easy to understand and its implementation to institutions' regulatory capital calculation process is relatively straightforward.

6.2.2. Generalized linear regression

Generalized linear regression has multiple specifications which can be used. GLM with gamma distribution is used in this study. The process for model development is very similar to linear regression model, expect that fitting a model to gamma distribution cannot handle zero values. To counter this restriction, observations with zero realized LGD are coded to value of 0.001. This is a minimal change as the newly coded value will be the lowest possible value, retaining risk characteristic similar. After this change the process follows familiar pattern: fit a model with all variables, use stepwise regression in variable reduction, and test variable combinations manually. Table 5 shows the results for the first model and the final variable combination. There is some variation in the final model compared to classical regression as more variables have statistically significant slopes.

Table 5. Initial generalized linear regression and the final model variable names, coefficients, and p values.

	Initial model		Final model	
Variable	Coefficients	P value	Coefficients	P value
Intercept	-2.188	0.0000	-2.190	0.0000
Amount of deposits	0.345	0.2638		
Business industry	1.093	0.0000	1.155	0.0000
Collateral type	0.427	0.0157	0.440	0.0119
Customer segment	0.125	0.5497		
Haircut value of collateral bundle	0.810	0.0000	0.823	0.0000
Issued amount	0.565	0.0001	0.553	0.0001
Number of debtors	0.601	0.0051	0.609	0.0042
Number of guarantees	0.041	0.8858	-0.442	0.0442
Number of pledges	-0.484	0.1252		
Percentage of paybacks	0.586	0.0021	0.571	0.0025
Time as customer	-0.054	0.8371		

Results of GLM model are very promising as the model can predict values from the whole scale of LGD. The highest estimated value is 0.82 with a mean of 0.12. Fitting to gamma distribution shows good result as observations are skewed around zero (Figure 3). The employed data does not have a bimodal distribution unlike some researchers have observed (Zhang & Thomas 2012). GLM model does not predict exact zero values as predictions are couple percentage points above zero. Regardless of this the predictions MAE is 0.168 and MSE is 0.071. Based on the measured estimation errors the GLM model performs very well. Pearson's correlation coefficient is 0.372 which shows prominent linear relationship between estimated and observed values.

In terms of R^2 of 0.135, the model leaves hope for better. Discriminatory power measured with AUC is 0.730. Based on these results GLM model with gamma distribution is a good method for LGD estimation. Model performs at least equally well with linear regression and even better based on some performance metric.

6.2.3. Beta regression

Beta regression modeling process follows same steps as other regressions. Process starts with all the variables in the data set and then variables are iteratively dropped or added to the model to find the best performing model. The final model has five explanatory variables and the intercept, all being statistically significant (Table 6). They are similar to other regressions with the exception to the number of pledges. The beta regression is defined to interval between zero and one both ends excluded. Due to this restriction, the dependent variable values outside the boundary are transformed to 0.001 and 0.999, respectively. This is a minimal change which keeps the observation riskiness and severity in right order as the transformed values are unique minimum and maximum for the realized LGD.

Table 6. Initial beta regression and the final model variable names, coefficients, and p values.

	Initial model		Final model	
Variable	Coefficients	P value	Coefficients	P value
Intercept	-1.365	0.0000	-1.369	0.0000
Amount of deposits	0.039	0.7718		
Business industry	0.203	0.0757		
Collateral type	0.227	0.0039	0.233	0.0029
Customer segment	0.076	0.3993		
Haircut value of collateral bundle	0.147	0.0194	0.141	0.0156
Issued amount	0.179	0.0038	0.183	0.0030
Number of debtors	0.047	0.6140		
Number of guarantees	0.002	0.9852		
Number of pledges	0.189	0.1730	0.201	0.0423
Percentage of paybacks	0.200	0.0154	0.192	0.0191
Time as customer	-0.073	0.5202		

Beta regression is studied by many past researchers but the results for this data set is not as outstanding as historical evidence suggests. The error metrics for predictions are 0.078 and 0.232 for MSE and MAE, respectively. Prediction errors, especially the absolute errors are bigger than those in linear regression. Also, the coefficient of determination shows a lack of explanatory power with a value of 0.039. While the R^2

is low, model's AUC is 0.726 which sets at the lower end of acceptable values. Lastly, correlation coefficient is 0.347 which is considered good. Based on these metrics, beta regression can be used to estimate LGD, but the data sets distribution may not match beta distribution thereby deteriorating error metrics.

The distribution of predicted values back-up the fact that beta regression is not the best model for the employed data set. Despite trying different variable and link functions, the model cannot predict the lowest and highest values of LGD. In fact, the predicted range is very narrow between 0.14 and 0.44. It may be that beta regression is better suited for more bimodally characteristic data implying high concentration around maximum values of LGD, as well.

6.2.4. Random forest

Random forest algorithm shows good prediction results. As the algorithm is based on fitting multiple regression trees the stepwise selection cannot be applied to variable selection. Instead, variable selection was done by manually trying different combinations. Chosen group of variables was in the end similar to other models. Seven variables were included and Table 7 summaries them. In tuning the algorithm, the number of trees of 500 was used. Two variables in the random sample of variables were tested at a time in each node as it was the best performing number.

Table 7. Variables used in the random forest model.

Variable name
Amount of deposits
Business industry
Collateral type
Haircut value of collateral bundle
Issued amount
Number of debtors
Percentage of paybacks

Prediction error metrics are low for random forest as MAE is 0.167 and MSE is 0.071. These low errors compare very well to other models. Predicted LGD values are from wide scale from 0.006 to 0.659. Low error might stem from the fact that the model is good in catching values very close zero as the median value of realized LGD is zero. Coefficient of determination is 0.133 which is in line with other models. In terms of discriminatory power random forest lacks a little. Its AUC value is 0.717. Overall, this method is good in LGD forecasting as the final metric, correlation is also quite high 0.368.

6.2.5. Support vector regression

Support vector regression model building is close to random forest. Stepwise regression does not apply in variable reduction. SVR is formed with all variables in it and then combinations with lesser number of variables are tested. The ultimate model was conducted based on the performance metrics preferring simpler model. Based on these, criteria variables are the same as in the final GLM model (Table 5).

Performance of SVR model is divided to clearly good and bad results. Good side of results is that MAE is as low as 0.138 and MSE is 0.083. One reason for low MAE may be the fact that predictions are weighted heavily to under 0.10 making the error in absolute terms low. This is due to the fact that major part of observed values are concentrated in the lower end of LGD range. Fragility of the model comes to light in the examination of other performance results. Coefficient of determination is negative with a value of -0.019. This is consequence of model’s residual sum of squares being more than total sum of squares e.g., model errors are bigger than with a prediction made using the average realized LGD.

Problematic results continue with AUC value of 0.666 and correlation of 0.262. Model’s discriminatory power to divide observations below and above the average LGD is below acceptable level of 0.70. Also, the correlation level is at the lower end of models. SVR performance can be summarized with one outstanding result, but other metrics show that the model is not consistently satisfying.

6.3. *Model performance comparison*

All five models are evaluated with the same performance metrics. The summary of all models with respective metric evaluation is shown in Table 8. It can be concluded that the best performing model is the GLM as it is best performing in terms of four metrics out of five. Support vector regression has the best MAE by far but in terms of all the other metrics it is also the worst by a clear distance. GLM's top performance is somewhat surprising as other models have been reported to have performed better in earlier research. Unexpectable results of GLM may stem from the unique distribution of realized LGD in the data set.

Table 8. Summary results from all models. Best individual metric in bold.

Model	MAE	MSE	R^2	AUC	Correlation
Generalized linear reg.	0.16841	**0.070562**	**0.135236**	**0.73967**	**0.37163**
Random forest	0.16725	0.070876	0.133170	0.71698	0.36793
Linear regression	0.17323	0.070966	0.132069	0.73043	0.36793
Beta regression	0.23197	0.078344	0.039242	0.72620	0.34742
Support vector reg.	**0.13787**	0.083311	-0.018930	0.66552	0.26194

The results show that generalized linear model, random forest, and linear regression stand out clearly from beta regression and support vector machine. The first of three models are very close to each other in terms of all metrics other than MAE. For example, MSE difference is only 0.0004 between GLM and linear regression, implying that GLM's ranking to be the best model is based on very narrow margin and no model can significantly outperform others. Performance of beta regression and SVR is surprisingly weak given the fact that both models have performed well in earlier papers. These results underline again the individualistic nature of IRB models as they are built specific to institutions own data and processes.

Overall ranking of the methods used was determined by giving one point to the best model in the respective metric and the worst model five points. This ranking method raises GLM as the best one, and the second-best method is random forest. These two models stand out over others as their performance is really similar. Other models are more distant from the top two. Ordering from the third to fifth place is linear regression, beta regression, and support vector regression.

6.4. *Conclusions from findings*

Previous research of LGD modeling has documented mixed results on the relative efficiency of estimation methodologies. The results of this thesis confirm the idea that there is no single best practice for the LGD modeling. The results show that linear model is not far from the level of non-linear models, which have performed better in more recent studies (Loterman et al. 2012). Somewhat surprise instead, the gamma GLM model performed best as it is used only in a few papers (Yaskir & Yaskir 2013). By contrary beta regression has been popularly used but it did not perform well in the test data set. Comparative results from this thesis underline the uniqueness of LGD models in terms of data used as some method may perform well on other data.

Figure 9 shows prediction errors for the best performing GLM model. Error is calculated as remainder of realized LGD and the predicted value. Thus positive values in the figure represent underestimation and negative values overestimation of LGD. It is clearly visible that overestimation is more common than underestimation. On the other hand underestimation looks to be more severe as the model cannot predict concistently high values of LGD. In fact the error is steadily increasing along with the realized LGD. On the top decile of realized LGD mean absolute error is 0.71 while for fifth decile, it is 0.44, and for the first -0.09. Prediction erros this big are a reasonable, although compared to a study by Loterman et al. (2012), they are significantly lower. It seems that LGD modeling needs more research to figure out the dependencies in order to improve estimation accuracy.

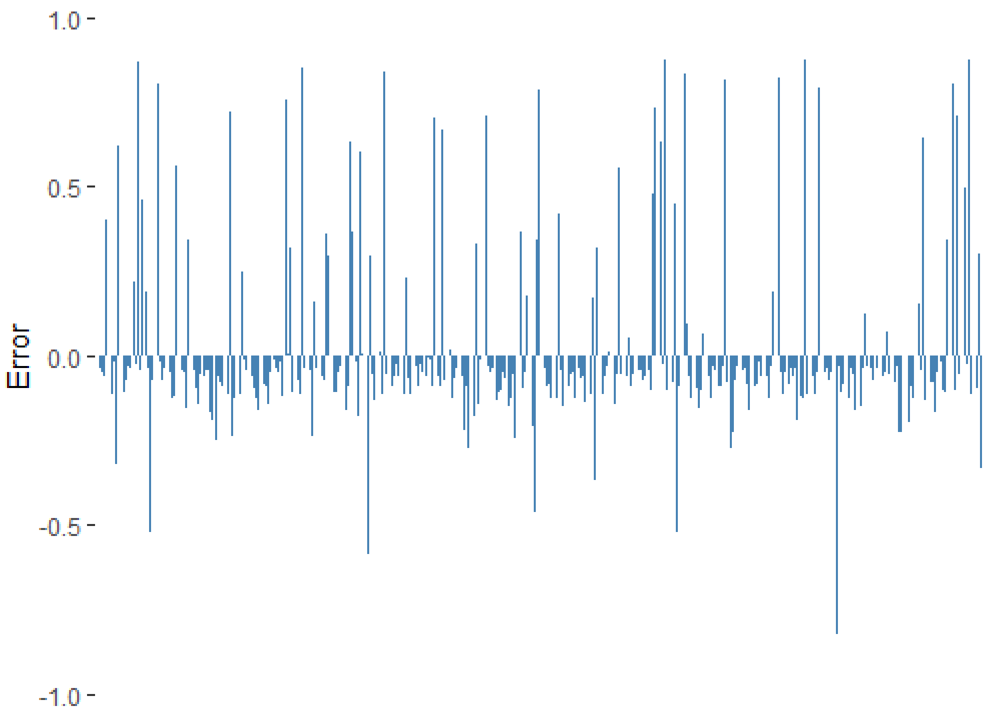


Figure 9. Prediction errors of the GLM model.

The predictive power between explanatory variables and realized LGD is mediocre at its best. This is no surprise as earlier research has noticed the same thing. Variables that performed better include collateral information e.g., type and value after haircut. Initially issued amount of debt and percentage of loan paid back were also among the significant explanatory variables, in line with the findings of Han & Jang (2013, 21). Conversely other end time as customer did not qualify to any of the models. Also the amount of bank deposits was unexpectedly a variable with low explanatory power. It may be that the information carried in the deposit amount is not perfect as debtors may have assets in other institutions which are not reflected in this variable.

7. Conclusion

This thesis studies LGD modelling from a perspective of retail credit risk management. Five different models were employed to forecast LGD in order to compare their performance. The models used were linear regression, generalized linear regression (gamma regression), beta regression, random forest, and support vector regression. Prediction performance was tested with MAE, MSE, coefficient of determination, AUC, and correlation coefficient. Overall ranking was formed by ranking models from the best performance to the worst with each of above-listed metrics and aggregating the rankings. The best performing model was generalized linear regression with gamma distribution in use. It was placed the best in all terms of metrics other than mean absolute error. The data set was WOE transformed with the technique introduced by Matuszyk et al. (2010) to test the method in LGD modeling as it is more popular in credit scoring.

LGD modeling is rather young topic as it was first introduced in Basel Committee for Banking Supervision's regulatory framework in 2004 (BCBS 2006). Earlier research has applied multiple modeling techniques on different types of loans (EBA 2017; Loterman et al. 2012). Still no consensus exists on the practicalities of LGD in contrast to other IRB risk-weight functions. This thesis contributes to the existing literature by focusing on retail credit as corporate credit has been more frequently studied topic. In addition, regulatory standards are discussed in greater detail from the perspective of EU legislation, which makes this study more practical compared to those referring only the Basel framework.

First of all, work-out LGD has been approved by EBA (2017) and should be used in LGD estimation. Work-out LGD is a technique where institutions own loss and recovery data is used from all defaulted obligations to build the LGD model. To comply with the regulation, institutions need to calculate realized LGD values of defaulted obligations in line with CRR (Regulation (EU) No 575/2013, art. 181) and EBA (16/2017, mom. 131-146).

For an institution to have an IRB model in use they need to acknowledge and comply with strict data requirements. CRR (Regulation (EU) No 575/2013) sets out standards for data scope, period, and risk drivers used in the modeling data set. Major requirements to take notice is the recording of all cashflows of all obligations until their final resolution. Institutions must ensure that recording of dates, amounts and types of cashflows continue from a year before default until the debt collection process is closed. EBA (16/2017, mom. 121) lists a comprehensive group of risk drivers which have to be included in the RDS or assessed in other way. A representative sample of risk drivers is used in the data set of this thesis.

One aspect of the study was to showcase the use of weight of evidence variable transformation in LGD modeling. WOE transformation has been used commonly in credit scorecard formation in measuring riskiness of loan (Thomas et al. 2005). To build on the work of Yap et al. (2011) and Matuszyk et al. (2010) in utilizing WOE transformation to LGD modeling, the modeling was conducted on transformed data. The process has many practical benefits. WOE transformed values are understandable, the calculation is simple to do, and it is computationally easy. The optimization of variable binning can be done with machine learning algorithm although in this study it was done manually. The transformation is most useful in missing value and outlier treatment. Binning process assigns missing values to unique bin and outliers do not have negative effect as they are dealt by binning them to maximum or minimum bin.

Earlier research has studied tens of methods to model LGD but as discussed no single method has outperformed others (Yashir & Yashir 2013; Loterman et al. 2012; Miller & Töws 2018). In this thesis five different models were employed and compared. The selection of the models was based on previous literature’s results. Among these five, the best-performing model is a generalized linear regression with a gamma distribution and logarithmic link function. The second best estimation accuracy was documented for the random forest approach which made it very close to gamma GLM model. The remaining three models were relatively far behind these two.

LGD models are very case sensitive as they are built on institutions internal data. Thus, data used to model LGD will vary depending on the institution’s business activities, strategy, and credit portfolio. This is the major reason why standard business practices have not been formed. LGD distribution is commonly skewed in bimodal or L-shape which supports standardization. Due to the skewness it can be misleading to assess average LGD. For these reasons, two-stage models have been used (Miller & Töws 2018). These models typically try first to separate low loss cases from high loss cases and then use other model to predict the actual LGD value. Although two-stage models may improve prediction accuracy the prediction of raw LGD was used in this thesis to better compare the methods being tested.

The resulting metrics show that non-linear models perform better than linear models. This is partly expected as the LGD is not normally distributed. Linear regression was used in the study as the baseline model as it is frequently used in LGD forecasting (Zhang & Thomas 2012). In the final ranking, linear regression was third best performing better than beta regression and support vector regression. In many models the forecasted values did not cover the whole range of LGD values. The lack of estimates in the higher end may be due to lack of observations, as 63 percent of observations resulted in zero LGD and only 12 percent in value higher than 0.5. With a larger data set a more robust model could be achieved.

Overall, all model’s performance metrics lack in quality. Resulting models do not have a good explanatory power in terms of with coefficient of determination. Even the best model has R^2 of 0.14 meaning that the models explain only 14 percent of the variance in the LGD values. On the other hand, coefficients of determination in the related research tend to be low in general as e.g. Loterman et al. (2012) compared 24 models and found low R^2 for all of them. In terms of mean absolute error, prediction values varied between 0.14 and 0.23. These errors are at acceptable level. Problems arise again because of the skewed data. Errors increase monotonically as observed LGD value increases. Further work is needed on finding a model that would be stable thorough all LGD values.

For further research topic I suggest data transformation in the context of LGD. Weight of evidence transformation was used in this thesis contrary to the major of earlier research of LGD. WOE transformation offers great benefits to the data processing thus saving resources as the process becomes more efficient. More research could be done in comparison of data transformation and standardization techniques for example, to build an LGD model on raw data, WOE transformed data, and transformed data with other techniques.

References

Altman, E. (1968) Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. The Journal of Finance (New York). 23 (4), 589–609.

Altman, E. et al. (2004) Default Recovery Rates in Credit Risk Modelling: A Review of the Literature and Empirical Evidence. Economic notes - Monte Paschi Siena. [Online] 33 (2), 183–208.

Altman, E. et al. (2005) The Link between Default and Recovery Rates: Theory, Empirical Evidence, and Implications. The Journal of business (Chicago, Ill.), 78(6), pp. 2203-2228.

Atkinson, P. & Blundell-Wignall, A. (2010) Thinking beyond Basel III: Necessary Solutions for Capital and Liquidity. OECD journal. Financial Market Trends. 2010 (1), 9–33.

Awad, M., & Khanna, R. (2015). Support vector regression. In Efficient learning machines (pp. 67-80). Apress, Berkeley, CA.

Bank of International Settlements (2021) History of Basel Committee. [www-document]. [Accessed 13 January 2021] Available https://www.bis.org/bcbs/history.htm

Bank of International Settlements (1988) International convergence of capital measurement and capital standards.

Basel Committee on Banking Supervision (2006) Basel II: Revised international capital framework.

Basel Committee on Banking Supervision (2011) Basel III: A global regulatory framework for more resilient banks and banking systems

Basel Committee on Banking Supervision (2017) High-level summary of Basel III reforms. [www-document]. [Accessed 7June, 2021] Available https://www.bis.org/bcbs/publ/d424_hlsummary.pdf

Basel Committee on Banking Supervision (2021) The Basel Framework. [www-document]. [Accessed 2 March, 2021] Available https://www.bis.org/basel_framework/index.htm?m=3%7C14%7C697

Black, F. & Scholes, M. (1973) The Pricing of Options and Corporate Liabilities. The Journal of Political Economy. 81 (3), 637–654.

Bluhm, C., Overbeck, L., & Wagner, C. (2016). Introduction to credit risk modeling. Crc Press.

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

Cribari-Neto, F., & Zeileis, A. (2010). Beta regression in R. Journal of statistical software, 34(1), 1-24.

Crook, J., Edelman, D. & Thomas, L. (2007) Recent developments in consumer credit risk assessment. European Journal of Operational Research. 183 (3), 1447–1465.

Draper, N. & Smith, H. (1998). Applied regression analysis. Wiley.

Duffie, D. & Singleton, K. (1999) Modeling Term Structures of Defaultable Bonds. The Review of Financial Studies. 12 (4), 687–720.

Dunn, P. K. & Smyth, G. K. (2018) Generalized Linear Models With Examples in R. New York, NY: Springer New York.

Enforcement Code (705/2007)

Eom, Y., Helwege, J., & Huang, J. (2004) Structural Models of Corporate Bond Pricing: An Empirical Analysis. The Review of Financial Studies. 17 (2), 499–544.

European Banking Authority (03/2016) Final Draft RTS on Assessment Methodology for IRB. [www-document]. [Accessed 15 March, 2021] Available https://www.eba.europa.eu/sites/default/documents/files/documents/10180/1525916/e8373cbc-cc4b-4dd9-83b5-93c9657a39f0/Final%20Draft%20RTS%20on%20Assessment%20Methodology%20for%20IRB.pdf?retry=1

European Banking Authority (07/2016) Guidelines on the application of the definition of default under Article 178 of Regulation (EU) No 575/2013 [www-document].

[Accessed 12 March, 2021] Available https://www.eba.europa.eu/regulation-and-policy/credit-risk/ guidelines-on-the-application-of-the-definition-of-default

European Banking Authority (2017) EBA Report on IRB modelling practices [www-document]. [Accessed 6 March, 2021] Available https://www.eba.europa.eu/ regulation-and-policy/model-validation/guidelines-on-pd-lgd-estimation-and-treatment-of-defaulted-assets

European Banking Authority (16/2017) Guidelines on PD estimation, LGD estimation and the treatment of defaulted exposures. [www-document]. [Accessed 18 May, 2021] Available https://www.eba.europa.eu/regulation-and-policy/model-validation/ guidelines-on-pd-lgd-estimation-and-treatment-of-defaulted-assets

European Banking Authority (2021) Implementing Basel III in Europe. [www-document]. [Accessed 1 February, 2021] Available https://www.eba.europa.eu/ regulation-and-policy/implementing-basel-iii-europe

European Commission (2021) Regulatory process in financial services. [www-document]. [Accessed 13 January, 2021] Available https://ec.europa.eu/info/business-economy-euro/banking-and-finance/regulatory-process-financial-services/regulatory-process-financial-services_en

European Commission (2021) Single supervisory mechanism. [www-document]. [Accessed 13 January, 2021] Available https://ec.europa.eu/info/business-economy-euro/banking-and-finance/banking-union/single-supervisory-mechanism_en#related links

Ferrari, S. & Cribari-Neto, F. (2004) Beta Regression for Modelling Rates and Proportions. Journal of Applied Statistics. 31 (7), 799–815.

Fight, A. (2004) Credit risk management. 1st edition. Oxford: Elsevier.

Financial Supervision Authority (2020) Capital adequacy and liquidity regulation (CRR/CRD). [www-document]. [Accessed 1 February, 2021] Available https://www.finanssivalvonta.fi/en/regulation/regulatory-framework/crrcrd/

Genuer, R. & Poggi, J.-M. (2020) Random Forests with R. Cham: Springer International Publishing AG.

Gürtler, M. & Hibbeln, M. (2013) Improvements in loss given default forecasts for bank loans. Journal of Banking & Finance. 37 (7), 2354–2366.

Han, C. & Jang, Y. (2013) Effects of debt collection practices on loss given default. Journal of Banking & Finance. 37 (1), 21–31.

Herring, R. (2018) The Evolving Complexity of Capital Regulation. Journal of Financial Services Research. 53 (2), 183–205.

Huajian Y. & Tkachenko, M. (2012) Modeling exposure at default and loss given default: empirical approaches and technical implementation. Journal of Credit Risk. 8 (2), 81–102.

Jarrow, R. & Turnbull, S. (2000) The intersection of market and credit risk. Journal of Banking & Finance. 24 (1), 271–299.

Jones, D. (2000) Emerging problems with the Basel Capital Accord: Regulatory capital arbitrage and related issues. Journal of Banking & Finance. 24 (1), 35–58.

Koch, S. et al. (2017) Bringing Basel IV into focus. McKinsey.

Koulafetis, P. (2017) Modern Credit Risk Management Theory and Practice. London: Palgrave Macmillan UK.

Leow, M. and Mues, C. (2012) Predicting loss given default (LGD) for residential mortgage loans: A two-stage model and empirical evidence for UK bank data. *International Journal of Forecasting,* 28(1), pp. 183-195.

Leymarie, J., Hurling, C. and Patin, A., (2018) Loss Functions for LGD Models Comparison. *European Journal of Operational Research,* 268(1), pp. 348-360.

Longstaff, F. & Schwartz, E. (1995) A Simple Approach to Valuing Risky Fixed and Floating Rate Debt. The Journal of Finance (New York). 50 (3), 789–819.

Loterman, G. et al. (2012) Benchmarking regression algorithms for loss given default modeling. International Journal of Forecasting. [Online] 28 (1), 161–170.

Matuszyk, A. et al. (2010) Modelling LGD for unsecured personal loans: decision tree approach. The Journal of the Operational Research Society. 61 (3), 393–398.

Merton, C. (1974) On the Pricing of Corporate Debt: The Risk Structure of Interest Rates. The Journal of Finance (New York). 29 (2), 449–470.

Miller, P. & Töws, E. (2018) Loss given default adjusted workout processes for leases. Journal of Banking & Finance. 91189–201.

Nguyen, Q. T. T. (2019) Basel III: where should we go from here? Journal of Financial Economic Policy. 11 (4), 457–469.

Olive, D. J. (2017) Linear Regression. Cham: Springer International Publishing.

Olson, D. & Wu, D. (2020) Predictive Data Mining Models. 2nd ed. 2020. Singapore: Springer Singapore.

Regulation (EU) (2013) No 575/2013 of the European Parliament and of the Council of 26 June 2013 on prudential requirements for credit institutions and investment firms and amending Regulation (EU) No 648/2012

Santos, J. A. C. (2001) Bank Capital Regulation in Contemporary Banking Theory: A Review of the Literature. Financial Markets, Institutions & Instruments. 10 (2), 41–84.

SAS Institute Inc. (2012) Building Loss Given Default Scorecard Using Weight of Evidence Bins in SAS® Enterprise Miner™ [www-document]. [Accessed 15 March, 2021] Available https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.421.1404&rep=rep1&type=pdf

SAS Institute Inc. (2013) Developing Credit Scorecards Using Credit Scoring for SAS® Enterprise Miner™ 13.1 [www-document]. [Accessed 15 March, 2021] Available https://documentation.sas.com/api/docsets/emcsgs/13.1/content/emcsgs.pdf?locale=en

Saunders, A. (2014) Financial institutions management. Macmillan Press.

Scandizzo, S. (2016) The validation of risk models: a handbook for practitioners. Palgrave Macmillan.

Schuermann, T. (2004). What do we know about loss given default? Federal Reserve Bank of New York.

Stupariu, P. et al. (2019) The disparity in PD and LGD estimates within the IRB framework and prospects for future improvement. Journal of Banking Regulation. 20 (4), 341–347.

Thomas, L., Oliver, R. & Hand, D. (2005) A survey of the issues in consumer credit modelling research. The Journal of the Operational Research Society. 56 (9), 1006–1015.

Win, S. (2018) What are the possible future research directions for bank's credit risk assessment research? A systematic review of literature. International Economics and Economic Policy. 15 (4), 743–759.

Witzany, J. (2017) Credit Risk Management Pricing, Measurement, and Modeling. Cham: Springer International Publishing.

Xie, S. (2021) Credit Risk Scorecard [www-document]. [Accessed 15 March, 2021] Available https://cran.r-project.org/web/packages/scorecard/index.html

Yan, X. & Su, X. (2009) Linear regression analysis theory and computing. Singapore: World Scientific.

Yap, B. et al. (2011) Using data mining to improve assessment of credit worthiness via credit scoring models. Expert Systems with Applications. 38 (10), 13274–13283.

Yashir, O. and Yashir, Y., 2013. Loss given default modeling: a comparative analysis. *Journal of Risk Model Validation,* **7**(1), pp. 25-59.

Zamore, S., Ohene Djan, l., Alon, I. & Hobdari, B. (2018) Credit Risk Research: Review and Agenda. Emerging Markets Finance & Trade. 54 (4), 811–835.

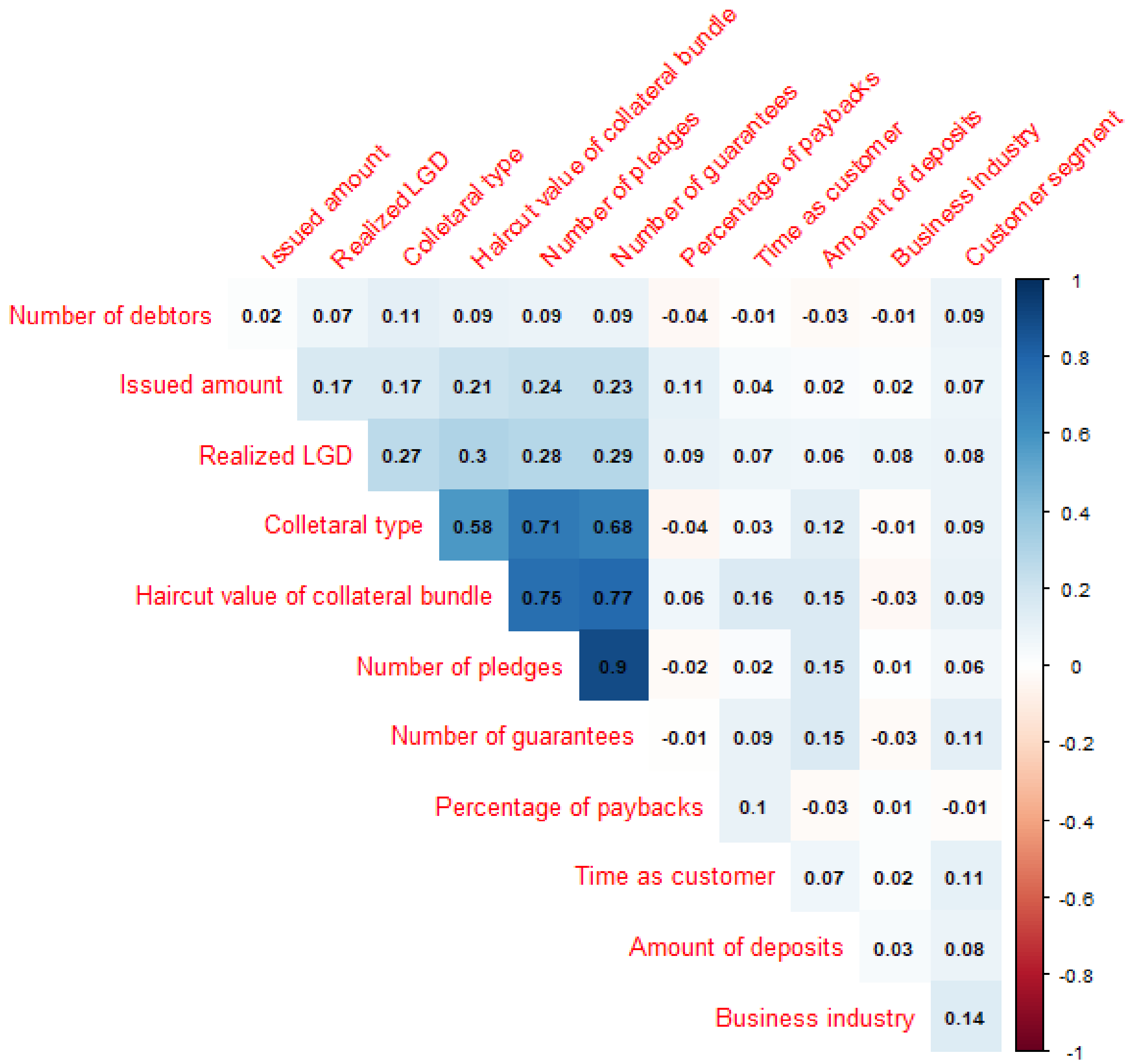
Zhang, J. & Thomas, L. (2012) Comparisons of linear regression and survival analysis using single and mixture distributions approaches in modelling LGD. International Journal of Forecasting. 28 (1), 204–215.

Appendix

Appendix 1. Capital requirement for retail exposure set out in Basel III.

$$Capital\ requirement = K = \left[LGD \cdot N\left[\frac{G(PD)}{\sqrt{(1-R)}} + \sqrt{\frac{R}{1-R}} \cdot G(0.999) \right] - PD \cdot LGD \right]$$

Appendix 2. WOE variable correlation matrix



	Issued amount	Realized LGD	Colletaral type	Haircut value of collateral bundle	Number of pledges	Number of guarantees	Percentage of paybacks	Time as customer	Amount of deposits	Business industry	Customer segment
Number of debtors	0.02	0.07	0.11	0.09	0.09	0.09	-0.04	-0.01	-0.03	-0.01	0.09
Issued amount		0.17	0.17	0.21	0.24	0.23	0.11	0.04	0.02	0.02	0.07
Realized LGD			0.27	0.3	0.28	0.29	0.09	0.07	0.06	0.08	0.08
Colletaral type				0.58	0.71	0.68	-0.04	0.03	0.12	-0.01	0.09
Haircut value of collateral bundle					0.75	0.77	0.06	0.16	0.15	-0.03	0.09
Number of pledges						0.9	-0.02	0.02	0.15	0.01	0.06
Number of guarantees							-0.01	0.09	0.15	-0.03	0.11
Percentage of paybacks								0.1	-0.03	0.01	-0.01
Time as customer									0.07	0.02	0.11
Amount of deposits										0.03	0.08
Business industry											0.14

Appendix 3. Explanatory variable clarification

Variable name	Explanation
Amount of deposits	Total amount of bank deposits
Business industry	Business industry of companies
Collateral type	Upper-level classification of collateral types
Customer segment	Segmentation of main livelihood
Haircut value of collateral bundle	Value of collateral after applying haircut
Issued amount	Initial amount of granted debt
Number of debtors	Number of joint debtors
Number of guarantees	Number of given guarantees
Number of pledges	Number of given pledges
Percentage of paybacks	Percentage of loan paid back
Time as customer	Time as customer measured in years