PREDICTING LOSS GIVEN DEFAULT

Gert Loterman

2013

Advisors:
Prof. dr. Geert Poels
Prof. dr. Manu De Backer

Dissertation submitted to the Faculty of Economics and Business Administration, Ghent University, in fulfillment of the requirements for the degree of Doctor in Applied Economics

Copyright

Doctoral committee

Prof. dr. Marc De Clercq
Dean-President, Ghent University

Prof. dr. Patrick Van Kenhove
Academic Secretary, Ghent University

Prof. dr. Geert Poels
Advisor, Ghent University

Prof. dr. Manu De Backer
Advisor, University College Ghent, Ghent University,
University of Antwerp

Prof. dr. Christophe Mues
University of Southampton

Prof. dr. Dirk Van den Poel
Ghent University

Prof. dr. Tony Bellotti
Imperial College London

0. DOCTORAL COMMITTEE

Acknowledgments

Knowledge is at the end based on acknowledgment. The gained knowledge during my five year PhD period, does not form an exception to this subtle statement. During this time I had the chance to gradually develop a scientific mindset which seems to me a very valuable asset. In addition, I also learned that developing interpersonal skills are equally important in work and life which do not always come spontaneous for an engineer. In what follows, I would like to thank the persons who supported and motivated me most during my PhD struggle.

I wish to thank, first and foremost, my advisors, prof. dr. Manu De Backer and prof. dr. Geert Poels. I very much appreciate the freedom they gave me to decide for myself when, where and how I worked on my research. They guided me through the entire PhD process which is not all fun and games. When I started, Manu mentioned the term 'doctoranditis'. I had no clue what he was talking about but now I understand the full meaning of this mental beast. Geert made sure I never deviated from the PhD path and guided me through all administrative hurdles. In particular the past half year, Geert was there to steer me towards the finish. Thank you Manu and Geert for supporting me - also in more difficult times.

Next to my advisors, I would like to thank my doctoral committee members, prof. dr. Christophe Mues, prof. dr. Dirk Van den Poel and prof. dr. Tony Bellotti. I sincerely owe much gratitude to Christophe who I consider to be my scientific mentor during my PhD term. Christophe was always there to think along, to carefully read my many draft texts and to provide me with numerous suggestions from the very beginning to the end. I look up to him. Many thanks also to both Dirk and Tony for critically proofreading the final PhD draft and for the useful comments and suggestions for improvement during the pre-defense.

Besides my doctoral committee, this thesis would not have been possible without the help of my co-authors. I would like to thank prof. dr. ir. David Martens, prof. dr. Christophe Mues and prof. dr. Bart Baesens to introduce me to the topic of LGD and to provide me with real-life LGD data from several international banks. In addition, I thank dr. Iain Brown for the collaboration on the benchmarking experiments and for his help on writing the first paper. I would also like to thank dr. Tony Van Gestel, dr. Karlien Vanden Branden and dr. Michiel Debruyne from Dexia Group to point to the literature gap in LGD backtesting and to provide me with ideas and real-life LGD data for further research on this subject.

Further, I would like to thank my teaching colleagues, Amy, Bart, Bertel, Damien, Els, Greet, Jan, Len and Manu, for providing me a

nice and motivating working environment. Special thanks of course to my fellow PhD candidates, Amy and Bart. Amy, thanks to double my - albeit small - fortune in Las Vegas. And yes, I fully agree with you on that monthly pastry event. Bart, I will not forget that you took the heaviest burden of our course on your shoulders, and so giving me the time to finish my PhD thesis. A word of thanks also to Martine who took care of all administrative hassles associated with finishing the PhD.

Finally, I owe much to my parents who supported me in every way throughout the PhD process. Nonetheless, I know they will be very relieved to know that my thirteen year old struggle of academic studies has finally come to an end. Thanks mom to take care of me and thanks dad for your everlasting motivational talks. Last but not least, thank you Veerle for walking with me the last years and to give me some nice perspectives in love and life. We plotted some fantastic plans together for the future and I will be very happy to execute these side by side.

I sincerely thank you all.

0. ACKNOWLEDGMENTS

Summary

The topic of credit risk modeling has arguably become more important than ever before given the recent financial turmoil. Conform the international Basel accords on banking supervision, financial institutions need to prove that they hold sufficient capital to protect themselves and the financial system against unforeseen losses caused by defaulters. In order to determine the required minimal capital, empirical models can be used to predict the loss given default (LGD). The main objectives of this doctoral thesis are to obtain new insights in how to develop and validate predictive LGD models through regression techniques.

The first part reveals how good real-life LGD can be predicted and which techniques are best. Its value is in particular in the use of default data from six major international financial institutions and the evaluation of twenty-four different regression techniques, making this the largest LGD benchmarking study so far. Nonetheless, it is found that the resulting models have limited predictive performance no matter what technique is employed, although non-linear techniques yield higher performances than traditional linear techniques. The results of this study strongly advocate the need for financial institutions to invest in the collection of more relevant

data.

The second part introduces a novel validation framework to backtest the predictive performance of LGD models. The proposed key idea is to assess the test performance relative to the performance during model development with statistical hypothesis tests based on commonly used LGD predictive performance metrics. The value of this framework comprises a solution to the lack of reference values to determine acceptable performance and to possible performance bias caused by too little data. This study offers financial institutions a practical tool to prove the validity of their LGD models and corresponding predictions as required by national regulators.

The third part uncovers whether the optimal regression technique can be selected based on typical characteristics of the data. Its value is especially in the use of the recently introduced concept of datasetoids which allows the generation of thousands of datasets representing real-life relations, thereby circumventing the scarcity problem of publicly available real-life datasets, making this the largest meta learning regression study so far. It is found that typical data based characteristics do not play any role in the performance of a technique. Nonetheless, it is proven that algorithm based characteristics are good drivers to select the optimal technique.

This thesis may be valuable for any financial institution implementing credit risk models to determine their minimal capital requirements compliant with the Basel accords. The new insights provided

in this thesis may support financial institutions to develop and validate their own LGD models. The results of the benchmarking and meta learning study can help financial institutions to select the appropriate regression technique to model their LGD portfolio's. In addition, the proposed backtesting framework, together with the benchmarking results can be employed to support the validation of the internally developed LGD models.

Summary (in Dutch)

Het topic kredietrisico modellering is, gezien de recente financiële crisis, misschien wel belangrijker dan ooit. Conform de internationale Basel akkoorden voor bankentoezicht dienen financiële instellingen aan te tonen dat ze over voldoende kapitaal beschikken om zichzelf en het financiële systeem te beschermen tegen onvoorziene verliezen veroorzaakt door wanbetalers. Om het vereiste minimale kapitaal te bepalen, kunnen empirische modellen worden gebruikt om het verlies bij wanbetaling of Loss Given Default (LGD) te voorspellen. De voornaamste doelstellingen van deze thesis zijn nieuwe inzichten te verkrijgen over hoe voorspellende LGD modellen te ontwikkelen en te valideren via regressietechnieken.

Het eerste deel laat zien hoe goed LGD kan worden voorspeld en welke technieken hiervoor het best zijn. De waarde zit in het bijzonder in het gebruiken van gegevens over wanbetalingen van zes grote internationale financiële instellingen en de evaluatie van vierentwintig verschillende regressietechnieken, wat dit de grootste LGD benchmarking studie maakt tot nu toe. Desalniettemin blijkt dat de resulterende modellen beperkt presteren ongeacht welke techniek gebruikt wordt, hoewel niet-lineaire technieken beter presteren dan traditionele lineaire technieken. De resultaten van deze studie tonen

sterk de noodzaak aan voor financiële instellingen om te investeren in het verzamelen van meer relevante gegevens.

Het tweede deel introduceert een nieuw backtesting raamwerk om de prestaties van LGD modellen te testen. Het voorgestelde sleutelidee is om de testprestaties te beoordelen ten opzichte van de prestaties tijdens de ontwikkeling van het model met statistische hypothesetesten gebaseerd op algemeen gebruikte metrieken voor het meten van de prestatie van LGD modellen. De waarde van dit raamwerk omvat een oplossing voor het gebrek aan referentiewaarden om te beslissen over al dan niet aanvaardbare prestaties en de vertekening van de prestaties door te weinig data. Dit onderzoek biedt financiële instellingen een praktisch instrument aan om de geldigheid van hun LGD modellen en bijbehorende voorspellingen te bewijzen zoals vereist door nationale toezichthouders.

Het derde deel onthult of de optimale regressietechniek kan worden geselecteerd op basis van de typische kenmerken van de data. De waarde zit vooral in het gebruik van het onlangs geïntroduceerde concept datasetoids die het genereren van duizenden datasets mogelijk maakt zodat het schaarsteprobleem van publiek beschikbare datasets kan verholpen worden, waardoor dit de grootste meta learning studie voor regressie tot dusver is. Het is gebleken dat typische data gebaseerde karakteristieken geen enkele rol spelen in de prestatie van een regressietechniek. Toch is het bewezen dat algoritme gebaseerde karakteristieken goede drijvers zijn om de meest optimale techniek te selecteren.

Deze thesis kan waardevol zijn voor elke financiële instelling die modellen implementeert voor kredietrisico om haar minimale kapitaalvereisten te bepalen en zo te voldoen aan de Basel akkoorden. De nieuwe inzichten in deze thesis kunnen een hulp bieden aan financiële instellingen om hun eigen LGD modellen te ontwikkelen en te valideren. De resultaten van de benchmarking en meta learning studie kunnen financiële instellingen helpen om de juiste regressietechniek te selecteren voor hun LGD portefeuilles. Daarnaast kan het voorgestelde backtesting raamwerk, samen met de benchmarking resultaten worden gebruikt om de validatie van de intern ontwikkelde LGD modellen te ondersteunen.

Contents

Copyright i

Doctoral committee iii

Acknowledgments v

Summary ix

Summary (in Dutch) xiii

1 Introduction 1

1.1 Overview 2

1.2 Literature review 5

1.2.1 Default 5

1.2.2 Loss 8

1.2.3 Prediction 11

1.3 Research goals 16

1.3.1 Problems 16

1.3.2 Questions 19

1.3.3 Methods 20

2 Benchmarking LGD models 23

2.1 Introduction 24
2.2 Literature review 25
2.3 Regression techniques 30
2.4 Performance metrics 42
2.5 Methods 46
2.5.1 Data collection 47
2.5.2 Algorithm configurations 49
2.5.3 Model evaluation 51
2.5.4 Implementation details 52
2.6 Results and discussion 52
2.7 Conclusions 58

3 Backtesting LGD models **59**
3.1 Introduction 60
3.2 Literature review 62
3.3 Proposed backtesting framework 67
3.3.1 Central tendency error tests 68
3.3.2 Dispersion error tests 70
3.3.3 Error, correlation and classification based tests 73
3.4 Methods 74
3.4.1 Data collection 75
3.4.2 Predictive modeling 77
3.4.3 Significance analysis 79
3.4.4 Power analysis 81
3.4.5 Implementation details 83
3.5 Results and discussion 83
3.6 Conclusions 88

4 Selecting LGD models **91**
4.1 Introduction . . . 92
4.2 Literature review . . . 96
4.2.1 Review of Rice's meta-learning framework . . 96
4.2.2 Review of previous meta-learning approaches . 98
4.3 Methods . . . 103
4.3.1 Algorithm space . . . 103
4.3.2 Performance mapping . . . 106
4.3.3 Feature mapping . . . 107
4.3.4 Problem space . . . 110
4.3.5 Selection mapping . . . 111
4.3.6 Meta model evaluation . . . 112
4.3.7 Implementation details . . . 113
4.4 Results and discussion . . . 113
4.5 Conclusions . . . 122

5 Conclusions **125**
5.1 Results . . . 126
5.2 Contributions . . . 128
5.3 Limitations . . . 130
5.4 Future research . . . 131

A Specification of the risk weight function **135**

B Results of the benchmarking experiment **147**

References **155**

1

Introduction

"If I owe you a pound, I have a problem;
but if I owe you a million, the problem is yours."
-John Keynes
(British economist, 1883-1946)

"Risk varies inversely with knowledge."
-Irving Fisher
(American economist, 1867-1947)

With the break out of the recent financial crisis, the topic of credit risk modeling has become more important than ever before. Financial institutions are investing heavily in the development of models to predict unforeseen losses in case debtors would fail to pay their obligations. It is crucial for banks to predict the potential loss of new loans in order to determine the minimal required capital to act as a safety cushion in case of defaults. The importance of research

in predicting Loss Given Default has nothing but strengthened because of the banking supervison by national regulators on Basel compliance. The goal of this study is to gain more insights in predicting Loss Given Default. This introductory chapter starts with a brief overview on how Loss Given Default drives a bank's minimal required capital conform with the Basel accords. Subsequently, a literature review is performed on regulatory requirements and predictive modeling of Loss Given Default. Based on the literature review, important literature gaps are identified and corresponding research proposals are outlined which are elaborated in the subsequent chapters.

1.1 Overview

The concept of banking dates back to the old Babylonian empire and, in essence, comes down on buying and selling financial products with corresponding profits and risks. A bank's main source of profits is generated through the difference between interests from lending activities and deposit interests. These activities go hand in hand with a certain risk that an obligor will default on its debt by failing to make payments which it is obliged to do. Since banks on their turn are buying from and selling to other banks, defaults can cause cascading failures and a collapse of an entire financial market. In order to protect the international financial system, international agreements are established in the Basel accords (1, 2, 3, 4). The accords aim to provide regulations to ensure that banks hold sufficient capital appropriate to the risks they are exposed to. Such capital can act as a safety cushion in case a sizeably larger proportion of

debtors default on their repayment obligations than provisioned for. The Basel accords are based on the principle that the required minimal capital to act as a safety cushion depends on the riskiness of a bank's assets. The more riskier a specific asset, the more capital is needed to absorb unexpected losses. Below is illustrated how banks may determine their minimal required capital according to the Basel regulations.

In the first Basel accord (1) a rather straightforward approach is suggested towards the calculation of the minimal required capital. It states that the ratio between the required capital and the value of the risk weighted asset should not be less than 8%, which is also known as the Cooke ratio:

$$\text{required capital} \geq 8\% \times \text{risk weight} \times \text{exposure}$$

where the risk weighted asset is the product of the exposure and the corresponding risk weight of a certain asset. Basel defines several risk categories to classify assets. Each category corresponds to a weight factor from 0% for extremely safe investments (e.g. sovereign debt) to 100% for very risky investments (e.g. corporate debt). For example, let's assume that a bank wants to cover a mortgage loan of 100000 EUR. Mortgage loans are labeled as moderately safe investments and represent a weight factor of 50%. If a bank tries to hold capital equal to 8% of its risk weighted assets, the minimal required capital will be:

$$\begin{aligned} \text{required capital} &\geq 8\% \times 50\% \times 100000 \text{ EUR} \\ &\geq 4000 \text{ EUR} \end{aligned}$$

This approach is fairly limited since it lacks nuances in risk weighting. Although assets in a particular risk category are labeled with the same corresponding risk weight, they do not always imply the same actual risks. For example, the risk weight of mortgage loans is 50% but the actual risk of a mortgage loan may be lower or higher depending on the amount of the obligor's monthly paycheck.

In order to quantify this actual risk more accurate, the second Basel accord (2) introduced the risk weight function where the required capital is driven by three key risk parameters to be estimated: PD the probability of default, LGD the loss given default and EAD the exposure at default.

$$\text{required capital} \geq f(PD) \times LGD \times EAD$$

where $f(.)$ is abstracted here for reasons of clarity but nonetheless further specified in Appendix A. In order to estimate these parameters for new loans, banks are encouraged to build internal models for each parameter based on their own historical loan data. For example, let's assume again that a bank wants to cover a mortgage of 100000 EUR and that internally built models estimate a PD of 3%, a LGD of 50% and an EAD of 90000 EUR (assume that 10000 EUR already is paid off at time of default), than the minimal required capital will yield:

$$\begin{aligned} \text{required capital} &\geq f(3\%) \times 50\% \times 90000 \text{ EUR} \\ &\geq 308 \text{ EUR} \end{aligned}$$

This approach is considered to be more risk sensitive as it takes into account varied factors which are empirically proven to be relevant in the bank's own data history. Note that this is known as the Internal Ratings Based (IRB) approach (5) and is also prevalent in the third Basel accord (3, 4).

1.2 Literature review

1.2.1 Default

Since LGD represents losses of defaulted issues, the definition of default is inherently connected to the LGD. There is a broad range of definitions of default, which can be classified as either subjective or objective (6). An objective definition is based on observable characteristics that are beyond the control of a bank (e.g. the grace period which represents the number of days past due). A subjective definition is based on risk managers appraisals or decisions made by the bank themselves (e.g. starting a legal process). The Basel definition of default is based on both a subjective and an objective condition. A default is defined as the occurrence when the obligor is past due more than ninety days on an obligation to the bank or when the bank considers that the obligor is unlikely to pay its obligation (§452 (2)). Note that the Basel definition of default applies

at the level of the obligor. In case of retail exposures, however, this can be applied at the level of a particular facility, rather than at the level of the obligor. Types of facilities are for example a loan or a bond. As such, defaults by a borrower on one obligation does not require a bank to treat all other obligations to the bank as defaulted (§455 (2)).

The first and objective part of the definition sets forth a grace period of ninety days. This is confirmed to be a good overall cut-off (7, 8). It is found that once obligors are ninety days in payment arrears, they remain in this delinquency status while only a minority recovers. A majority of the obligors with less than ninety days in payment arrears are most likely to recover. Hence, this point-of-no-return justifies the Basel grace period of ninety days. Note that Basel allows supervisors to define a default after a grace period of 180 days instead of ninety days, in case of retail and PSE (Public Sector Entity) exposures, if appropriate to local conditions (§452 (2)). The second and subjective part of the definition incorporates the unlikeliness-to-pay formulation so as to give supervisors a certain degree of freedom to take into account particularities of their jurisdiction (9). The meaning of unlikeliness to pay is clarified in Basel as a series of six elements, i.e. the bank puts the obligation in non-accrued status, the bank makes a charge-off, the bank sells the obligation with loss, the bank consents to a distressed restructuring of the obligation, the bank files for obligor's bankruptcy, the obligor is placed in bankruptcy (§453 (2)). Supervisors have to provide appropriate guidance as to how these elements must be implemented

and monitored (§454 (2)).

Banks often do not have sufficient data from defaulted facilities (e.g. a portfolio of large corporate loans) to estimate LGD. Therefore, they might consider the use of external data sources (e.g. rating agencies or pooled data across institutions). If external estimates of LGD are based on another definition of default, it is necessary to adjust for the difference in the definition of default (6). An LGD dataset consists of defaulted issues only. As a consequence, the definition of default defines the LGD. Basel allows banks to use external default data if the differences in the default definition are carefully analyzed and made consistent (§462 (2)). Therefore, it is useful to develop methods to establish a link between LGD estimates which use different default definitions (10). Rating agencies (e.g. Moody's, Fitch, S&P) apply their own default definitions (11, 12, 13). These may differ on how they treat missed payments that were made during a grace period or missed payments because of commercial disputes (7). Additionally, the grace period applied by rating agencies varies compared to Basel, see Table 1.1.

Definition	**Grace period**
Basel	90 - 180 days
Moody's	0 days
Fitch	10 - 30 days
S&P	10 - 30 days

Table 1.1: Comparison of grace period (7)

1.2.2 Loss

Basel defines LGD as the economic loss expressed as a percentage of the exposure in case of default (§297 (2)). It is important to notice that the economic loss (i.e. real loss) as defined by Basel is not the same as the accounting loss (i.e. bookkeeping loss) (2, 6, 7). The economic loss must include material discount effects and material direct and indirect costs associated with collecting on the exposure (§460 (2)). To calculate the economic loss using the observed recoveries and costs, it is necessary to discount them back to the date of default using some discount rate. The impact of the chosen discount rate is particularly important in portfolio's where the recovery period is long and has a low risk level (14, 15, 16). Direct costs are those associated with a particular asset (e.g. a fee for an appraisal of collateral). Indirect costs are necessary to carry out the recovery process but are not associated with individual facilities (e.g. overhead associated with the office space for the workout department).

The LGD can be measured via subjective methods or objective methods. Subjective methods on the one hand are based on qualitative expert judgment. These are particularly used for portfolios with no or few defaults. Objective methods on the other hand are based on quantitative information about the economic loss. Objective methods can be subdivided into either explicit or implicit methods, see Table 1.2. Explicit methods on the one hand use the market value (market LGD) or discounted cash-flows from the recovery process (workout LGD) from defaulted facilities to determine the LGD. Implicit methods on the other hand derive the LGD

Source	Measure	Methods	Exposure
Market values	Price differences	Market LGD	Large corporate, sovereigns, banks
	Credit spreads	Implied market LGD	Large corporate, sovereigns, banks
Recovery and cost experience	Discounted cash flows	Workout LGD	Retail, SMEs, large corporate
	Historical losses and estimated PD	Implied historical LGD	Retail

Table 1.2: Classification of the objective methods to obtain LGDs (6)

from the expected loss from the credit spread of risky bonds (implied market LGD) or the historical total losses (implied historical LGD) and the probability of default of non-defaulted facilities. The method to be employed depends on the exposure as illustrated in Table 1.2.

Workout LGD and implied historical LGD is driven by the recovery and cost experience of the exposure. Workout LGD is calculated by discounting cash flows and costs, resulting from the workout from the date of default to the end of the recovery process. Both cash and non-cash recoveries as well as direct and indirect costs have to be determined as accurately as possible. In addition, it is important to use an appropriate discount factor which is the subject of considerable disagreement amongst practitioners and banking supervisors (15, 16, 17). Further, banks must define when a workout is finished. Sometimes banks employ a recovery threshold (e.g. when the remaining non-recovered value is lower than 5% of the EAD) or a given time threshold (e.g. one year from the date of default) (6).

The implied market LGD is determined by looking at the credit spreads of the non-defaulted risky bonds. The credit spread reflects the expected loss on the bonds next to a liquidity premium (18). Recent models illustrate how to decompose this measure of expected loss into the PD and the LGD (19, 20). Because the implied market method uses information from non-defaulted facilities, there is some debate whether this method is valid from a regulatory perspective (6).

Market LGD and implied market LGD is driven by the market value of the exposure. Market LGD is computed by comparing the face value of a facility before default and the market value of the facility after time of default. The price difference is a measure for the economic loss, expressed as a percentage of the exposure (i.e. the face value of the facility). The rating agency (21, 22, 23) recovery studies are based on this approach and typically evaluate the market value of the defaulted facility about thirty days after default (7, 18). The market prices reflect the discounted expected recovery and thus implicitly represent the economic loss. However, if markets are driven by fluctuations unrelated to the expected recovery, this measure may not be appropriate (6). Implied historical LGD is obtained from the estimate of the PD and the experience of total losses in the portfolio (§465 (2)). Consequently, the LGD can than be determined according to the formula: Expected Loss (EL) = Probability of Default (PD) x Loss Given Default (LGD). This method may be useful for retail exposures because in most cases it is easier to estimate the PD than the LGD (7).

1.2.3 Prediction

According to the Basel accords, LGD estimates must be grounded in historical experience and empirical evidence (§465 (2)). The most common technique to meet this requirement is to build an LGD model through regression analysis of historical default data. The resulting LGD model can subsequently be used for predicting unknown LGD values for new customers. Regression analysis allows to determine the relationship between a number of potential drivers of LGD (i.e. the independent variables) and the LGD on the other hand (i.e. dependent variable) based on a dataset of defaulted borrowers. Numerous techniques exist to perform regression analysis. For a detailed overview of regression algorithms for LGD modeling is referred to Chapter 2. Note that the Basel accords require the data observation period to build an LGD model to be minimal five years for retail exposures (§473 (2)) and minimal seven years for corporates, sovereigns and bank exposures (§472 (2)). Hence, it is ensured that empirically build LGD models cover at least one complete economic cycle of default behavior.

When building a predictive LGD model it is of crucial importance both to obtain correct outcomes and to understand how an LGD model comes to its conclusions. Therefore, an LGD model is required to be both accurate and comprehensible. A model is said to be accurate when the difference between its predicted values and the observed values is small. The observed or realized LGD is the ex

post measure of the realized economic loss, expressed as a percentage of the exposure at time of default while the predicted or expected LGD is the ex ante estimate of the economic loss conditional on the default (6). Note that the realized LGD is also the complement of the Recovery Rate (RR), i.e. LGD = 1 - RR. The accuracy is most often measured by quantifying the similarity between predictions and observations. Various performance metrics to measure model accuracy are described in Chapter 2. A model is said to be comprehensive when the relation between the LGD and its drivers can be well interpreted and explained by a human being. Although not straightforward to measure, the degree of comprehensibility depends on the complexity of the type of model output. While the accuracy measures the data fit, the comprehensibility measures the mental fit of the model (24).

The Basel accords require banks to estimate LGD to reflect economic downturn conditions where necessary to capture the relevant risks (§468 (2)). The LGD may be lower in periods of recession and the estimated LGD should be conservative enough in order not to underestimate the actual loss. Two modeling approaches can be distinguished in order to capture stressed economic conditions. One way is to take into account cyclical effects in order to reflect economic downturn conditions where necessary. Cyclical effects might be captured by including macro-economic factors in the predictive model (25). However, when the model fails in capturing downturn conditions, LGD may be underestimated and can cause high losses. Another way is not to take into account cyclical effects but instead

to rely upon an overly conservative LGD in order to capture the relevant risk in periods of economic downturn (26). A drawback, however, is an overestimated LGD in general. This may needlessly increase the capital requirements and hence may cause banks to be less competitive. Note that the Basel accords require that the LGD cannot be less than the long term average Loss Given Default calculated based on the average economic loss of all observed defaults within the data source for that type of facility (§468 (2)).

The identification of the most important drivers of LGD is crucial for building high predictable models. Commonly used variables for LGD analysis can be classified in features of the issuer, features of the issue, macroeconomic factors and the relation between bank and borrower (7). First, the features of the counterpart include the creditworthiness of the borrower, the industry sector classification and industry conditions, the size, the legal structure, age, country of residence and its legal environment, balance-sheet structure, financial flexibility to increase revenues to repay debt in case of distress, number of creditors. Second, the features of the issue are characterized by absolute and relative seniority, product type, type and value of the collateral, guarantees, exposure/size, length and costs of the workout process, maturity and syndication. Third, macroeconomic factors include economic conditions, default rate levels, interest rate levels, gross domestic product, growth, etc. Forth, the relation between bank and borrower is important such as intensity of the relation of the bank with the counterpart, length of the relation.

Commonly used variables to characterize retail customers, corporates and local governments are listed below (7). For retail customers (27, 27, 28, 29, 30, 31), typical application scoring variables are sociodemographic variables, financial indicators, product information and customer information and typical behavioral score variables are flow variables, interval measures, customer relation measures, product status management, flash volume variables, debt level and debt burden and demographic customer information. For firm counterparts (32, 33, 34, 35, 36) income statement and balance sheet information allows construction of typical quantitative variables such as profitability, leverage and gearing, growth, liquidity, activity, size and volatility. Although banks and insurance companies are closely related to firms, it is important for these counterparts to measure the size of the equity buffer with respect to the risks the insurer or bank are exposed to. Typical variables for local governments (37) are debt, exploitation, self-financing ability, macroeconomic and demographic elements and size.

In order to characterize insurance companies, banks and sovereigns, the following variables are most frequently used (7). For insurance companies (38), typical variables are capital adequacy, leverage and debt, performance and profitability, liquidity, cash flow and size. The variables for banks (39, 40, 41, 42, 43, 44) are typically organized along the CAMEL variables, i.e. capital adequacy, asset quality, management, earnings and liquidity. Financial information on countries and sovereigns is available from official international sources like the IMF and the World Bank. Typical variables

for sovereigns are social development level, macroeconomic environment, debt, states and markets, state efficiency, stability, political regime. Important differences of sovereigns and public sector entities with firm counterparts are legal and institutional differences. Although macroeconomic and demographic variables are important as well, the health of the public sector entity is determined by the strength of the local economy and the management of the local authority.

The most important LGD drivers appear to be the security and priority of the claims according to a series of empirical studies (45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55). Secured debt and high priority decrease the LGD. Another important driver turns out to be the default rate (47, 48, 49, 54, 55, 56, 57). LGD is typically higher in a period of high defaults. Other macroeconomic variables did not seem to matter when the default rate was taken into account (45, 47). The industry sector (45, 46, 57) and the liquidity of the collateral (52, 58) also seems to affect LGD. Industry sectors with credits that are backed up by liquid collateral (e.g. cash or accounts receivable) seem to experience a lower LGD than industry sectors backed up by less liquid collateral (e.g. property or equipment). Further, the size of the borrower (51) and the size of the loan (46, 50, 51, 59) did not tend to affect LGD. Note that the majority of these studies cover corporate LGD.

According to the Basel studies on the validation of internal rating systems (6), several main open issues in the area of LGD validation

require further research. A first open issue is about how to determine realized LGD. The Basel report demonstrates the importance of several decisions which may affect LGD estimation. These include dealing with negative losses, choosing the interest rate for discounting losses and deciding when the recovery process is over. A second open issue is about which estimation techniques are most appropriate for LGD estimation. The Basel report highlighted that the use of simple techniques such as averages may be misleading given the typical non-normal distribution of LGD. Hence, further research is needed on more advanced regression analysis of rich LGD datasets including multiple risk drivers. A third open issue is about how to compare different LGD models (i.e. benchmarking) and how to compare realized and estimated LGD (i.e. backtesting). Although these validation procedures are regulatory requirements, the Basel report is not explicit on which techniques to use for this purpose.

1.3 Research goals

1.3.1 Problems

First, the current empirical LGD literature is not clear about which regression models may fit real-life LGD best. Although credit risk modeling research has largely focused on the estimation of the PD parameter (6, 60), the LGD parameter may have a larger impact on capital requirements. The latter enters the Basel risk weight function in a linear way, unlike PD which has less of a direct effect on minimal required capital. Hence, any changes in the LGD

model estimates have a strong bearing on the capital of a financial institution and as such also its long-term strategy. It is thus of crucial importance to have models that estimate LGD as accurate as possible. This seems however not a trivial issue as the empirical LGD literature typically reports low performances and does not agree which regression technique is best suited for LGD modeling. Suggested models are often built using simple averages (61, 62), (generalized) linear regression (25, 61, 62, 63, 64, 65, 66, 67, 68) or regression trees (25, 61, 64). The low accuracy results may be caused either by the use of limited regression techniques or data with limited predictability. Up to now, no empirical LGD study in the literature has focused on gaining more insights in this matter by assessing different state-of-the-art techniques on a multitude of different LGD datasets.

Second, the current literature is not clear on how to validate internal LGD models. Basel requires financial institutions to regularly validate its internal estimation process and its internal models but does not mention how this may be done (6). The assessment of a model's predictions typically includes backtesting which is the process of evaluating to which degree the internal LGD model estimates correspond with the realized LGD observations. Commonly used performance metrics in the empirical LGD literature include MSE (25, 62, 64), RMSE (61, 63, 69), MAE (61, 63, 64), R^2 (25, 65, 67, 68) and AUROC (62, 66, 68). It is however not straightforward to determine acceptable accuracy solely based on these metrics. After all, a single value has little meaning without

an appropriate reference value indicating acceptable accuracy. In addition, these metrics do not take into account the number of LGD observations. When the portfolio lacks sufficient observations, a few extreme observations can distort the accuracy result and so degrade its reliability. Recent research has largely focused on backtesting PD models (70, 71, 72) while literature on statistical hypothesis testing for LGD models is non-existing.

Third, the current literature is not clear whether and how dataset characteristics may drive the fitting performance of regression models in general or LGD models in particular. In order to build a model to fit the typical non-normal characteristics of LGD data better, many studies suggest to transform the LGD prior to linear regression. These result in models such as tobit models (25, 64), logit models (25, 66), logistic models (61, 63, 65, 68), log-log models (61, 63, 66, 67) or beta models (25, 62, 66, 69). Nonetheless, it is not proven that these significantly fit LGD data better. Apart from LGD studies, many meta-learning studies claim that commonly used dataset characteristics (e.g. size, dimensionality, composition, distribution, landmarks) may favor a specific predictive model algorithm (73, 74, 75, 76, 77). However, the lack of sufficient real-life datasets available to these meta-learning studies (i.e. merely twenty-two (77) to hundred (78)) undermine the support of these claims. In spite of the arsenal on meta-learning studies, it is not clear how commonly used dataset characteristics drive regression algorithm fitting performance.

1.3.2 Questions

Research Question 1: How accurate can regression models fit real-life LGD?

The objective is to uncover to which degree regression techniques can fit a model to real-life LGD data. The predictive power of real-life LGD should be clearly quantified by fitting various algorithms to various LGD datasets. Additionally, any statistically significant performance differences between regression techniques should be quantified. Both the identification of the independent variables which drive the observed real-life LGD and the detailed relationships between these drivers and the LGD is however out of scope.

Research Question 2: How can the predictive performance of LGD models be evaluated?

The objective is to develop a framework of tests in order to allow financial institutions to support the validation of their internal LGD models. The tests should be applied in such a way that they can determine upon acceptable model performance. The tests should be able to detect when the accuracy of an LGD model is significantly deteriorating. In addition, the tests should take into account the influence of possible accuracy distortion caused by a possible lack of sufficient observations. The study of the evaluation of the LGD model performance in low default portfolios are beyond the scope.

Research Question 3: How can dataset characteristics drive the

fitting performance of regression models?

The objective is to gain insight whether and which dataset characteristics drive the fitting performance of regression algorithms. Any additional value of a meta model based on data characteristics or algorithm based characteristics compared to a meta model with simple training averages should be clearly quantified and statistically tested. Data based characteristics involve the number of instances, dichotomous variables, continuous variables and distribution properties of the dependent variable while algorithm based characteristics involve the algorithm's performance on very small data samples. Since LGD models are required to be comprehensible, only algorithms which lead to a humanly understandable output form are part of the scope.

1.3.3 Methods

The first research question is answered by applying the framework on the statistical comparison of classifiers over multiple datasets by Demsar (79) and its extensions by Garcia and Herrera (80). The experiments are based on real-life LGD datasets which are obtained from six international financial institutions, each of which contains data about defaulted loans and their resulting losses. The types of loan portfolios included are personal loans, corporate loans, revolving credit and mortgage loans. A varied arsenal of both most commonly used regression techniques and performance metrics to fit the real-life LGD datasets and to assess the model fit respec-

tively, is employed. In total, eight performance metrics are employed to assess twenty-four regression techniques applied on six real-life LGD datasets from major international banks. The averaged performances are statistically compared using Friedman's test (81) and the post-hoc multiple testing procedure of Hommel (82) to detect any significant differences between every regression technique and the best performing one. More details and results are discussed in Chapter 2. Note that this study is also published as 'Loterman et al. (2012). Benchmarking regression algorithms for loss given default modeling. International Journal of Forecasting, 28: 161-170.'

The second research question is answered by proposing a workbench of statistical hypothesis tests for LGD backtesting, analogous to a recently introduced PD backtesting framework (70). The proposed workbench includes standard parametric tests (i.e. T-test and F-test) (83), standard non-parametric tests (i.e. Wilcoxon signed rank test and Ansari-Bradley test) (84, 85) and a number of non-standard tests constructed through a bootstrapping approach based on commonly used performance metrics in LGD literature (i.e. RMSE, MAE, AUROC, AOREC, R^2, r, ρ and τ) (86, 87, 88, 89). These tests are applied in such a way that they take into account an appropriate reference value indicating acceptable accuracy in addition to the number of LGD observations. The proposed backtesting framework is demonstrated on a linear model based on real-life LGD data which reflects corporate loan loss rates over a time span from 1984 to 2004 and contains 891 observations. Further, all tests are subject to a statistical power analysis in order to evaluate the reliability of

the proposed tests. More details and results are discussed in Chapter 3.

The third research question is answered by applying the framework on the algorithm selection problem by Rice (90). The experimental data is constructed by implementing the recently introduced concept of datasetoids (91, 92) on the algorithm selection problem. A datasetoid is defined as a new dataset obtained by switching an independent variable with a dependent variable. This idea allows to circumvent the scarcity of publicly available real-life datasets (93) by generating more than thousand regression datasetoids to build up a meta dataset. The meta dataset consist of dataset characteristics as independent variables and the performance differences of the considered algorithms as dependent variables. In the context of LGD analysis, the experiments involve comprehensible regression models only (i.e. linear, spline, tree, linear tree and spline tree). Both a data characteristics based meta model and an algorithm characteristics based meta model is statistically compared with each other and a simple training average based meta model using Friedman's test (81) followed by the Holm post-hoc pairwise testing procedure (94) to determine any significant performance differences (79, 80). More details and results are discussed in Chapter 4.

2

Benchmarking LGD models

"Prediction is very difficult, especially if it's about the future."
-Niels Bohr
(Danish physicist, 1885-1962)

"Study the past if you would devine the future."
-Confucius
(Chinese philosopher, 551-479 BCE)

In this large-scale LGD benchmarking study, various regression techniques to model and predict LGD are investigated. These include one-stage models, such as those built by ordinary least squares regression, beta regression, robust regression, ridge regression, regression splines, neural networks, support vector machines and regression trees, as well as two-stage models which combine multiple techniques. In total 24 techniques are compared using six real-life loss datasets from major international banks. It is found that much of

the variance in LGD remains unexplained as the average prediction performance of the models in terms of R^2 ranges from 4% to 43%. Nonetheless, a clear trend can be observed that non-linear techniques and in particular support vector machines and neural networks perform significantly better than more traditional linear techniques. Also, two-stage models built by a combination of linear and non-linear techniques are shown to have similarly good predictive power, while they offer the added advantage of having a comprehensible linear model component.

2.1 Introduction

Credit risk research has so far largely focused on the estimation and validation of the PD parameter, i.e. the likelihood of a default. The LGD parameter on the other hand measures the economic loss, expressed as a percentage of the exposure, in case of default. In other words, LGD is the proportion of the remaining loan amount that the bank would not be able to recover. This parameter is a crucial input to the regulatory capital calculations as it enters the Basel risk weight function in a linear way (unlike PD, which therefore has less of a direct effect on minimal capital). Hence, any changes in the LGD estimates produced by models have a strong bearing on the capital of a financial institution and as such also its long-term strategy.

It is thus of crucial importance to have models that estimate LGD as accurately as possible. This seems however not a trivial issue as the empirical LGD literature typically reports low performances.

Such models are often built using simple averages, (generalized) linear regression or regression trees. The low accuracy results may be caused either by the use of limited regression techniques or data with limited predictability. Up to now, no empirical LGD study in the literature has focussed on gaining more insights in this matter by assessing different techniques on a multitude of different LGD datasets. This first large scale LGD benchmarking study investigates using a set of six real-life default loss datasets whether other approaches can improve the prediction performance of these LGD models.

The remainder of this chapter is organized as follows. First, a literature review is conducted on empirical studies which explicitly focus on modeling LGD for the purpose of forecasting. Second, an overview is given of both the examined regression techniques and the performance metrics used to evaluate and compare the models. Third, the available real-life LGD datasets are described and the experimental set up is outlined in order to perform the benchmarking experiments. Forth, the obtained experimental results are reported and discussed and are followed with a conclusion.

2.2 Literature review

The literature on empirical studies which focus on forecasting LGD is rather limited. Since LGD estimation has not been a regulatory requirement since the advent of the second Basel accord, few institutions do not have a sufficiently large LGD dataset at the moment to build and validate a predictive LGD model. In addition, banks

are not eager to share these for scientific research because of reasons of confidentiality if they do have a large track record of losses. The select amount of empirical studies on forecasting LGD employ datasets which are mainly American (62), Portugese (61, 67), German (95), Italian (64, 65), British (25) or Czech (66). The largest LGD dataset in terms of time span covers more than three decades of default losses and dates back to as early as 1981 (62). The datasets vary in size from as small as 374 defaults (61) to as large as 134937 defaults (64). These include portfolio's such as loans to SMEs (61, 64, 65, 66, 67, 95), large corporate loans (62, 66), credit card accounts (25) and personal loans (64, 65, 95).

Based on the literature, the LGD distribution is typically non-normal distributed but most often rather bimodally distributed. Real life LGD tends to be characterized by high concentrations of either total recovery or total loss or both. The majority of the empirical literature studies report of a large peak on zero and a smaller peak on one (61, 62, 64, 66, 67). Caselli et al. (65) report the opposite: a large peak on one and a smaller peak on zero. Bellotti and Crook (25) even observe equally large peaks on both zero and one for credit card accounts. Nonetheless, Gurtler and Hibbelz (95) observe only a large peak on zero while Gupton (62) observe only a large peak on one for the corporate loan segment. Similar observations are obtained in LGD studies which do not focus on forecasting LGD (49, 50, 68, 96, 97). Based on these studies, there does not seem to be an obvious connection between the relative size of the peaks on zero and one and the type of portfolio. Note that these

may be caused by factors as internal bank policies or external economic conditions.

The most basic modeling practice observed in the empirical literature is the use of a simple historical average which often functions as a benchmark to compare the predictive performance of more advanced techniques (61, 62). Various regression techniques are employed in the emperical literature to model real-life LGD. The most often used technique seems to be ordinary least squares which builds linear models (25, 64, 65, 66, 68). Given the bimodal distribution of LGD which violates the normality assumption of ordinary least squares, several alternatives are proposed to circumvent this issue. These result in models such as tobit models (25, 64), logit models (25, 66), logistic models (61, 63, 65, 68), log-log models (61, 63, 66, 67) and beta models (25, 62, 66, 69). Further, experiments are also done with non-linear techniques such as regression trees (25, 61, 64) and neural networks (63).

The evaluation of the predictive performance of a model in the empirical literature is generally done by comparing the LGD model predictions with the actual realized LGD dataset observations. These may be error based such as the Mean Squared Error (MSE) (25, 62, 64), the Root Mean Squared Error (RMSE) (61, 63, 69) or the Mean Absolute Error (MAE) (61, 63, 64). Although not used in an LGD context, Bi and Bennet (86) proposed an alternative error based metric, i.e. the Area Above the Regression Error Characteristics curve (AOREC), which could also be used to assess the predictive

performance of LGD models. Other metrics observed in the LGD literature are correlation based such as Pearson's product-moment correlation coefficient r (62), Kendall's correlation coefficient τ (66) or the Coefficient of Determination R^2 (25, 65, 67, 68). Note that Spearman's rank correlation coefficient ρ (98) could be an alternative performance metric. Finally, even classification based metrics are proposed to assess the predictive performance of LGD models such as the Area Under the Receiver Operation Characteristics curve (62, 66, 68).

Different techniques are compared with each other in the literature using the above mentioned methods. Based on these studies, it is not clear which technique is best for LGD predictive modeling. Gupton et al. (62) found that their LossCalc model based on beta regression is more accurate that models based on a simple historical average based on 3026 defaulted corporate loans and bonds. According to Calabrese et al. (64) a (joint) beta regression model is better than a linear, tobit or decision tree model based on experiments on 134937 defaulted loans to SMEs. Bellotti and Crook (25) on the other hand report that linear models are better than beta, tobit, logit and decision tree models based on 55000 defaulted credit card accounts. Bastos et al. (61) reports that decision trees appear to be more accurate than historical averages, log-log and logistic models based on 374 defaulted loans granted to SMEs. According to Chalupta et al. (66) logit models appear to be slightly better than linear, log-log and beta models on a few hundred defaulted corporate and SME loans.

Based on previous empirical studies, LGD models typically show weak predictive performance. Gupton et al. (62) recorded a performance of 0.42 to 0.68 in terms of Pearson's r and of 0.70 to 0.80 in terms of AUROC on 3026 defaulted corporate loans and bonds. Dermine et al. (67) reported an R^2 performance of 0.08 to 0.20 based on 10000 loans to SMEs. Bellotti and Crook (25) obtained similar R^2 results ranging from 0.01 to 0.20 based on 55000 defaulted credit card accounts. Chalupta et al. (66) reported a performance of 0.38 to 0.42 in terms of Kendall's τ and of 0.58 to 0.66 in terms of AUROC based on a few hundred corporate and SME loans. Gurtler et al. (95) obtained relatively higher R^2 results of 0.25 to 0.60 and an average AUROC of 0.73 based on 69985 defaulted personal loans and loans to SMEs. Casselli et al. (65) reported similar R^2 results of 0.42 to 0.66.

In order to make well founded conclusions about the predictive performance of regression model techniques, a statistical evaluation on multiple datasets is required, which is lacking in the empirical LGD literature. For this purpose, Demsar (79) provided a workbench of statistical hypothesis tests in order to detect significant differences between techniques in terms of predictive performance. In first instance, it is suggested to use the Friedman's test (81) in order to statistically test the null hypothesis that there is no difference between the multiple hold-out validation performance of the techniques on multiple datasets. When this null hypothesis can be statistically rejected, it is suggested to use a pairwise post-hoc test-

ing procedure to statistically test the null hypothesis that a pair of techniques differ in multiple hold-out validation performance, e.g. Nemenyi test (99). Garcia and Herrera (80) extended the workbench of Demsar with more powerful pairwise post-hoc tests, e.g. Hommel test (82).

2.3 Regression techniques

This is an overview of the regression techniques for the benchmarking experiments. These include the most popular techniques found in the empirical LGD literature supplemented with more advanced machine learning techniques commonly applied for regression tasks in general. Both one stage and two stage techniques are considered. One stage techniques can be divided into linear and nonlinear techniques. Linear techniques model the dependent variable as a linear function of the independent variables while nonlinear techniques fit a nonlinear model to a dataset. Two stage models are a strategic combination of the aforementioned one stage models. These either combine the comprehensibility of an OLS model with the added predictive power of a non-linear technique, or they use one model to first discriminate between zero and higher LGDs and a second model to estimate LGD for the subpopulation of nonzero LGDs.

The following mathematical notations are employed to describe the techniques in a more formal way. A scalar x is denoted in normal script. A vector $\mathbf{x}$ is represented in boldface and is assumed to be a column vector. The corresponding row vector $\mathbf{x}^T$ is obtained

using the transpose T. Bold capital notation is used for a matrix $\mathbf{X}$. The number of independent variables is given by n and the number of observations is given by l. The observation i is denoted as $\mathbf{x}_i$ whereas variable j is indicated as x_j. The value of variable j for observation i is represented as $x_i(j)$ and the independent variable y for observation i is represented as y_i. P is used to denote a probability. A regression technique fits a dataset to a model $y = f(\mathbf{x}) + e$ where y is the dependent variable, $\mathbf{x}$ are the independent variables and e is the residual.

Ordinary Least Squares (OLS)

Ordinary least squares regression (87) is the most common technique to find optimal parameters $\mathbf{b}^T = [b_0\ b_1\ b_2\ ...\ b_n]$ to fit a linear model to a dataset as

$$y = \mathbf{b}^T\mathbf{x}$$

where $\mathbf{x}^T = [1\ x_1\ x_2\ ...\ x_n]$. OLS approaches this problem by minimizing the sum of squared residuals:

$$\sum_{i=1}^{l}(e_i)^2 = \sum_{i=1}^{l}\left(y_i - \mathbf{b}^T\mathbf{x}_i\right)^2$$

By taking the derivative of this expression and subsequently setting the derivative equal to zero

$$\sum_{i=1}^{l}(y_i - \mathbf{b}^T\mathbf{x}_i)\mathbf{x}_i^T = \mathbf{0}$$

the model parameters $\mathbf{b}$ can be retrieved as

$$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

with $\mathbf{X}^T = [\mathbf{x}_1\ \mathbf{x}_2\ ...\ \mathbf{x}_l]$ and $\mathbf{y} = [y_1\ y_2\ ...\ y_l]^T$.

Ridge Regression (RiR)

Ridge regression (100) is a linear regression variant that is less sensitive to correlated independent variables than OLS. When independent variables are strongly correlated with each other, inverting the $\mathbf{X}^T\mathbf{X}$ matrix leads to large and unreliable parameter estimates. Ridge regression reduces these undesirable symptoms by minimizing

$$\lambda \mathbf{b}^T\mathbf{b} + \sum_{i=1}^{l}(e_i)^2 = \lambda \mathbf{b}^T\mathbf{b} + \sum_{i=1}^{l}(y_i - \mathbf{b}^T\mathbf{x}_i)^2$$

where λ is defined as the ridge parameter which controls a trade-off between bias and variance. With values of λ larger than zero, the model parameters are more biased but can be estimated more reliably as

$$\mathbf{b} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$$

where $\mathbf{I}$ is the identity matrix.

Robust Regression (RoR)

Robust regression (101) is another linear regression variant that is less sensitive to outliers as OLS. When the dataset contains outliers, the model parameters can become unreliable. Therefore, the most common method for robust regression called M-estimation (102) minimizes

$$\sum_{i=1}^{l}\rho(e_i) = \sum_{i=1}^{l}\rho(y_i - \mathbf{b}^T\mathbf{x}_i)$$

where the objective function $\rho(e)$ should be less sensitive for outliers than the function used by OLS, i.e. $\rho(e) = e^2$. By taking the

derivative of the objective function and subsequently setting the derivative equal to zero

$$\sum_{i=1}^{l} w_i(y_i - \mathbf{b}^T\mathbf{x}_i)\mathbf{x}_i^T = \mathbf{0}$$

where $w(e) = \dfrac{\dfrac{\partial\rho}{\partial e}}{e}$ is defined as the weight function and $w_i = w(e_i)$ are the resulting weights. Because the weights depend upon the residuals, the residuals depend upon the estimated coefficients and the estimated coefficients depend upon the weights, the solution requires an iterative procedure (Iteratively Reweighted Least Squares or IRLS). To start, the initial model parameters $\mathbf{b}^{(0)}$ are estimated by setting $w_i = 1$ as in OLS. At each iteration t, the model parameters $\mathbf{b}^{(t)}$ are estimated using the residuals $e_i^{(t-1)}$ and associated weights $w_i^{(t-1)}$ from the previous iteration. The new estimates are given by

$$\mathbf{b}^{(t)} = (\mathbf{X}^T\mathbf{W}^{(t-1)}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}^{(t-1)}\mathbf{y}$$

where $\mathbf{W}^{(t-1)} = diag\left\{w_i^{(t-1)}\right\}$. This procedure stops when the estimated model parameters $\mathbf{b}$ satisfy a convergence criterion (103).

Ordinary Least Squares with Beta transformation (B-OLS)

Whereas OLS regression tests generally assume normality of the dependent variable y, the empirical distribution of LGD can often be approximated more accurately by a Beta distribution (104). Assuming that y is constrained to the open interval $(0, 1)$, the cumulative distribution function (CDF) of a Beta distribution is given by:

$$\beta(y; a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^y v^{a-1}(1-v)^{b-1}dv$$

where $\Gamma()$ denotes the well-known Gamma function, and a and b are two shape parameters, which can be estimated from the sample mean μ and variance σ^2 using the method of the moments, i.e.:

$$a = \frac{\mu^2(1-\mu)}{\sigma^2} - \mu\,;\;\; b = a(\frac{1}{\mu} - 1)$$

A potential solution to improve model fit therefore is to estimate an OLS model for a transformed dependent variable $y_i^* = N^{-1}(\beta(y_i; a, b))$ $(i = 1, ..., l)$, in which $N^{-1}()$ denotes the inverse of the standard normal CDF. The predictions by the OLS model are then transformed back through the standard normal CDF and the inverse of the fitted Beta CDF to get the actual LGD estimates.

Beta Regression (BR)

Instead of performing a Beta transformation prior to fitting an OLS model, an alternative Beta regression model approach can be considered (105). This model for estimating a dependent variable bounded between zero and one is closely related to the class of generalized linear models and allows for a dependent variable that is Beta-distributed conditional on the covariates. Instead of the usual parametrization though of the Beta distribution, with shape parameters a and b, they propose an alternative parametrization involving a location parameter μ and a precision parameter ϕ, by letting:

$$\mu = \frac{a}{a+b}\,;\;\; \phi = a + b$$

It can be easily shown that the first parameter is indeed the mean of a $\beta(a, b)$-distributed variable, whereas $\sigma^2 = \frac{\mu(1-\mu)}{(\phi+1)}$, so for fixed μ, the variance (dispersion) increases with smaller ϕ.

Two link functions mapping the unbounded input space of the linear predictor into the required value range for both parameters are then chosen, viz. the logit link function for the location parameter (as its value must be squeezed into the open unit interval) and a log function for the precision parameter (which must be strictly positive), resulting in the following sub models:

$$\mu_i = E(y_i|\mathbf{x}_i) = \frac{e^{\mathbf{b}^T\mathbf{x}_i}}{1 + e^{\mathbf{b}^T\mathbf{x}_i}}$$

$$\phi_i = e^{-\mathbf{d}^T\mathbf{x}_i}$$

This particular parametrization offers the advantage of producing more intuitive variable coefficients (as the two rows of coefficients, $\mathbf{b}^T$ and $\mathbf{d}^T$, provide an indication of the effect on the estimate itself and its precision, respectively). By further selecting which variables to include in (or exclude from) the second submodel, one can explicitly model heteroskedasticity. The resulting log-likelihood function is then used to compute maximum-likelihood estimators for all model parameters.

Ordinary Least Squares with Box-Cox transformation (BC-OLS)

The aim of the family of Box-Cox transformations (106) is to make the residuals of the regression model more homoskedastic and closer to a normal distribution. The Box-Cox transformation on the dependent variable y_i takes the form

$$\begin{cases} \dfrac{((y_i + c)^\lambda - 1)}{\lambda} & \text{if } \lambda \neq 0 \\ log(y_i + c) & \text{if } \lambda = 0 \end{cases}$$

with power parameter λ and parameter c. If needed, the value of c can be set to a non-zero value to rescale y so that it becomes strictly positive. After a model is built on the transformed dependent variable using OLS, the predicted values can be transformed back to their original value range.

Regression trees (RT)

Classification and regression trees are decision tree models, for a categorical or continuous dependent variable, respectively, that recursively partition the original learning sample into smaller subsamples, so that some impurity criterion $i()$ for the resulting node segments is reduced (107). To grow the tree, one typically uses a greedy algorithm that, at each node t, evaluates a large set of candidate variable splits so as to find the 'best' split, i.e. the split s that maximizes the weighted decrease in impurity:

$$\Delta i(s,t) = i(t) - p_L i(t_L) - p_R i(t_R)$$

where p_L and p_R denote the proportions of observations associated with node t that are sent to the left child node t_L or right child node t_R, respectively. A commonly applied impurity measure $i(t)$ for regression trees is the mean squared error or variance for the subset of observations falling into node t. Alternatively, a split may be chosen based on the p-value of an ANOVA F-test comparing between-sample variances against within-sample variances for the subsamples associated with its respective child nodes (ProbF criterion).

Multivariate Adaptive Regression Splines (MARS)

MARS (108) is a technique that uses piecewise linear functions to capture non-linearities and interactions between variables. The method is based on a 'divide and conquer' strategy where the input space is divided in partitions and each partition holds its own regression equation. MARS fits a dataset to a model of the form

$$y = \sum_{k=1}^{K} b_k B_k(\mathbf{x}) + e$$

where $B(\mathbf{x})$ is a basis function and K refers to the number of basis functions. A basis function can either take the value one or a single hinge function $h(x_j)$ that takes the form of $max(0, x_j - a)$ or $max(0, a - x_j)$ with a a so-called knot, or a product of 2 or more hinge functions to model interactions. MARS builds a model in 2 phases: a forward and a backward pass. The forward pass builds an over fitted model by adding a number of Hinge functions, typically twice the number of Hinge functions with the lowest mean squared error. Both variables and knots are selected via a partition scheme and a subsequent exhaustive search. The backward procedure prunes the model by removing those Hinge functions that are associated with the smallest increase in the so-called GCV (Generalized Cross Validation) error, defined as

$$GCV = \frac{\sum_{i=1}^{l}(y_i - f(\mathbf{x}_i))^2}{(1 - \frac{C}{l})^2}$$

where $C = 1 + c \cdot d$, c is a penalty for adding a Hinge function and d is the number of independent Hinge functions.

Least Squares Support Vector Machines (LSSVM)

In this study an SVM (109) variant, called LSSVM (110), is used because of its higher efficiency for solving large scale problems (111). The basic idea behind regression with LSSVM is to map the independent variables to a high dimensional feature space with a non-linear function φ so the data becomes more appropriate for linear regression:

$$y = \mathbf{b}^T \varphi(\mathbf{x}) + e$$

with $\varphi^T(\mathbf{x}) = [1\ \varphi(x_1)\ \varphi(x_2)\ ...\ \varphi(x_n)]$. However, the model is never evaluated in this form. Instead, LSSVM regression fits a model to a dataset by minimizing

$$\frac{1}{2}\mathbf{b}^T\mathbf{b} + \frac{1}{2}\gamma \sum_{i=1}^{l}(e_i)^2 = \frac{1}{2}\mathbf{b}^T\mathbf{b} + \frac{1}{2}\gamma \sum_{i=1}^{l}(y_i - \mathbf{b}^T\varphi(\mathbf{x}_i))^2$$

where γ is defined as the regularization parameter. The primal optimization problem indicates that each data point has to be mapped to a high dimensional (possibly infinite) feature space. This mapping however becomes quite fast computationally infeasible. To bypass this problem, the kernel trick is used. In order to be able to do the kernel trick, the optimization problem has to be reformulated in its dual form by applying the method of Lagrange multipliers that leads to the following equation:

$$y = \sum_{i=1}^{l} \alpha_i \varphi(\mathbf{x})^T \varphi(\mathbf{x}_i) + e$$

At this point the kernel trick can be performed. The kernel K is a function that calculates the dot products of the input vectors in

feature space without implicitly doing the mapping to the feature space. The kernel trick is supported by Mercer's theorem and replaces every dot product in high dimensional feature space by a simple kernel function:

$$K(\mathbf{x}, \mathbf{x}_i) = \varphi(\mathbf{x})^T \varphi(\mathbf{x}_i)$$

Artificial Neural Networks (ANN)

ANNs are mathematical representations inspired by the functioning of the human brain (112). The benefit of an ANN is its flexibility in modeling virtually any (non-linear) dependency between independent variables and the dependent variable. Although various architectures have been proposed, our study focuses on probably the most widely used type of ANN, i.e. the Multilayer Perceptron (MLP). A MLP is typically composed of an input layer (consisting of neurons for all input variables), a hidden layer (consisting of any number of hidden neurons), and an output layer (in our case, one neuron). A common way of training ANNs is backpropagation. Each neuron processes its inputs and transmits its output value to the neurons in the subsequent layer. Each such connection between neurons is assigned a weight during training. The output of hidden neuron i is then computed by applying an activation function $f^{(1)}$ to the weighted inputs and its bias term $b_i^{(1)}$ (having a similar role to the intercept of a regression model) as follows:

$$h_i = f^{(1)}(b_i^{(1)} + \sum_{j=1}^{n} \mathbf{W}_{ij} x_j)$$

$\mathbf{W}$ is the weight matrix whereby $\mathbf{W}_{ij}$ denotes the weight connecting input j to hidden neuron i. Similarly, the output of the output layer

is computed as follows:

$$y = f^{(2)}(b^{(2)} + \sum_{j=1}^{n_h} \mathbf{v}_j h_j)$$

with n_h the number of hidden neurons and $\mathbf{v}$ the weight vector whereby $\mathbf{v}_j$ represents the weight connecting hidden neuron j to the output neuron. Examples of transfer functions that are commonly used are the sigmoid function $f(x) = \frac{1}{1+e^{-x}}$, the hyperbolic tangent $f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$ and the linear transfer function $f(x) = x$.

During model estimation, the weights of the network are first randomly initialized and then iteratively adjusted so as to minimize an objective function, typically the sum of squared errors (possibly accompanied by a regularization term to prevent over fitting). This iterative procedure can be based on simple gradient descent learning or more sophisticated optimization methods such as Levenberg-Marquardt or Quasi-Newton. The number of hidden neurons can be determined through a grid search based on validation set performance.

Linear regression + non-linear regression (OLS+)

The purpose of this two-stage technique is to combine the good comprehensibility of OLS with the predictive power of a non-linear regression technique (113). In a first stage, a linear model

$$y = \mathbf{b}^T \mathbf{x} + e$$

is built with OLS. In a second stage, the residuals e of this linear model

$$e = g(\mathbf{x}) + e^*$$

are estimated with a non-linear regression model g in order to further improve the predictive ability of the model. Doing so, the model takes the following form:

$$y = \mathbf{b}^T\mathbf{x} + g(\mathbf{x}) + e^*$$

where e^* are the new residuals of estimating e. A combination of OLS with RT, MARS, LSSVM and ANN is assessed in this study.

Logistic regression + (non)linear regression (LOG+)

The LGD distribution is often characterized by a large peak around LGD = 0. This non-normal distribution can lead to inaccurate regression models. This proposed two-stage technique attempts to resolve this issue by modeling the peak separately from the rest. Therefore, the first stage of this two-stage model consists of a logistic regression to estimate whether LGD ≤ 0 or LGD > 0. In a second stage the mean of the observed values of the peak is used as prediction in the first case and a one-stage (non)linear regression technique is used as prediction in the second case. More specifically, a logistic regression (114) results in an estimate of the probability P of being in the peak

$$P = \frac{1}{1 + e^{-(\mathbf{b}^T\mathbf{x})}}$$

with $(1 - P)$ as the probability of not being in the peak. This two-stage model is built using the following equation:

$$y = P \cdot \overline{y}_{peak} + (1 - P) \cdot f(\mathbf{x}) + e$$

where $\overline{y}_{peak}$ is the mean of the values of y ≤ 0, which practically equals to 0, and $f(\mathbf{x})$ is a one-stage (non)linear regression model, build on those observations only that are not in the peak. Whereas $\overline{y}_{peak}$ is determined using only the values of y ≤ 0, the one-stage model is built using only the values of y > 0. A combination of logistic regression with all aforementioned one-stage techniques as described above, is assessed is this study.

2.4 Performance metrics

This is an overview of the performance metrics to evaluate the extend to which degree regression model predictions $f(\mathbf{x}_i)$ differ from the dataset observations y_i of the dependent variable. These include the most popular performance metrics found in the empirical LGD literature supplemented with performance metrics applied for regression tasks in general. Each of these metrics has its own method of quantifying model performance.

Root Mean Squared Error (RMSE)

RMSE is defined as the square root of the average of the squared difference between predictions and observations:

$$RMSE = \sqrt{\frac{1}{l}\sum_{i=1}^{l}(f(\mathbf{x}_i) - y_i)^2}$$

RMSE has the same units as the dependent variable being predicted. Since residuals are squared, this metric heavily weights outliers.

The RMSE is bound between the maximum squared error and zero (perfect prediction).

Mean Absolute Error (MAE)

MAE is given by the averaged absolute differences of predicted and observed values:

$$MAE = \frac{1}{l}\sum_{i=1}^{l}|f(\mathbf{x}_i) - y_i|$$

Just like RMSE, MAE has the same unit scale as the dependent variable being predicted. Unlike RMSE, MAE is not that sensitive to outliers. The metric is bound between the maximum absolute error and zero (perfect prediction).

Area under the Receiver Operating Characteristic Curves (AUC)

ROC curves are normally used for the assessment of binary classification techniques (89). It is however used in this context to measure how good the regression technique is in distinguishing high values from low values of the dependent variable. To build the ROC curve, the observed values are first classified into high and low classes using the mean $\overline{y}$ of the training set as reference. The area under the ROC curve (AUC) is an estimate for the discriminatory power of the technique. The AUROC varies from 0.5 (random classification) to one (perfect classification).

Area over the Regression Error Characteristic curves (AOC)

REC curves (86) generalize ROC curves for regression. The AOC curve plots the error tolerance on the x-axis versus the percentage of points predicted within the tolerance (or accuracy) on the y-axis. The resulting curve estimates the cumulative distribution function of the squared error. The area over the REC curve (AOC) is an estimate of the predictive power of the technique. Unlike the AUROC, the AOREC is bound between zero (perfect prediction) and the maximum squared error.

Coefficient of Determination (R^2)

The Coefficient of Determination R^2 (87) can be defined as one minus the fraction of the residual sum of squares to the total sum of squares:

$$R^2 = 1 - \frac{SS_{err}}{SS_{tot}}$$

where $SS_{err} = \sum_{i=1}^{l}(y_i - f(\mathbf{x}_i))^2$, $SS_{tot} = \sum_{i=1}^{l}(y_i - \overline{y})^2$ and $\overline{y}$ is the mean of the observed values. Since the second term in the formula can be seen as the fraction of unexplained variance, the R^2 can be interpreted as the fraction of explained variance. The R^2 is usually expressed as a number on a scale from zero to one. However, R^2 can yield negative values when the model predictions are worse than using the mean $\overline{y}$ from the training set as prediction.

Pearson's Correlation Coefficient (r)

Pearson's r (98) is defined as the sum of the products of the standard scores of the observed and predicted values divided by the degrees of freedom:

$$r = \frac{1}{l-1} \sum_{i=1}^{l} \left(\frac{y_i - \overline{y}}{s_y} \right) \left(\frac{f(\mathbf{x}_i) - \overline{f})}{s_f} \right)$$

with $\overline{y}$ and $\overline{f}$ the mean and s_y and s_f the standard deviation of respectively the observations and predictions. Pearson's r can take values between minus one (perfect negative correlation) and one (perfect positive correlation) with zero meaning no correlation at all.

Spearman's Correlation Coefficient (ρ)

Spearman's ρ (98) is defined as Pearson's r applied to the rankings of predicted and observed values. If there are no or few tied ranks however, it is more usual to use the equivalent formula

$$\rho = 1 - \frac{6 \sum_{i=1}^{l} d_i^2}{l(l^2 - 1)}$$

where d_k is the difference between the ranks of observed and predicted values. Spearman's ρ can take values between minus one (perfect negative correlation) and one (perfect positive correlation) with zero meaning no correlation at all.

Kendall's Correlation Coefficient (τ)

Kendall's τ (98) measures the degree of correspondence between observed and predicted values. In other words, it measures the association of cross tabulations:

$$\tau = \frac{n_c - n_d}{\frac{1}{2}l(l-1)}$$

where n_c is the number of concordant pairs and n_d is the number of discordant pairs. A pair of observations $\{i, k\}$ is said to be concordant when there is no tie in either observed or predicted LGD (i.e. $y_i \neq y_k$, $f(\mathbf{x}_i) \neq f(\mathbf{x}_k)$), and if $sgn(f(\mathbf{x}_k) - f(\mathbf{x}_i)) = sgn(y_k - y_i)$, where i, $k = 1, ..., l$ $(i \neq k)$. Similarly, it is said to be discordant if there is no tie and if $sgn(f(\mathbf{x}_k) - f(\mathbf{x}_i)) = -sgn(y_k - y_i)$. Kendall's τ can take values between minus one (perfect negative correlation) and one (perfect positive correlation) with zero meaning no correlation at all.

2.5 Methods

This section describes the collected real-life LGD datasets and outlines the experimental benchmarking framework used to assess the performance of the various models built on the real-life LGD datasets. After data pre-processing, the models are built on the training sets and predictive performance metrics are reported for the remaining test sets. Several of the included techniques require parameter settings or tuning and/or benefit from variable selection; further details of both are provided below, along with the procedure used to assess whether the observed performance differences are statistically

significant.

2.5.1 Data collection

The LGD datasets are obtained from six financial institutions, each of which contains loan-level data about defaulted loans and their resulting losses. An overview of the datasets is given in Table 2.1. The number of dataset entries varies from a few thousands to just under 120000 observations. The number of available input variables ranges from twelve to forth-four. The types of loan portfolios included are personal loans, corporate loans, revolving credit and mortgage loans. The empirical distribution of LGD values observed in each of the datasets is displayed in Figure 2.1. Note that the LGD distribution in consumer lending often contains one or two spikes around LGD = 0 (in which case there was a full recovery) and/or LGD = 1 (no recovery). Also, a number of datasets include some LGD values that are negative (e.g., because of penalties paid, gains in collateral sales, etc.) or larger than one (e.g., due to additional collection costs incurred); in other datasets, values outside the unit interval were truncated to zero or one by the banks themselves. Importantly, in none of these datasets, LGD appears to be normally distributed. More information on these datasets is confidential.

Prior to the benchmarking experiments, the datasets are pre-processed as follows. Instances with missing values are excluded from the dataset. Each dataset is randomly shuffled and divided into two-thirds training set and one-third test set. The training set is used to build the models while the test set is used solely to assess the pre-

Dataset	Type	Inputs	Total size	Training size	Test size
BANK1	Personal loans	44	47853	31905	15948
BANK2	Mortgage loans	18	119211	79479	39732
BANK3	Mortgage loans	14	3351	2232	1119
BANK4	Revolving credit	12	7889	5260	2629
BANK5	Mortgage loans	35	4097	2733	1364
BANK6	Corporate loans	21	4276	2851	1425

Table 2.1: Overview of dataset characteristics

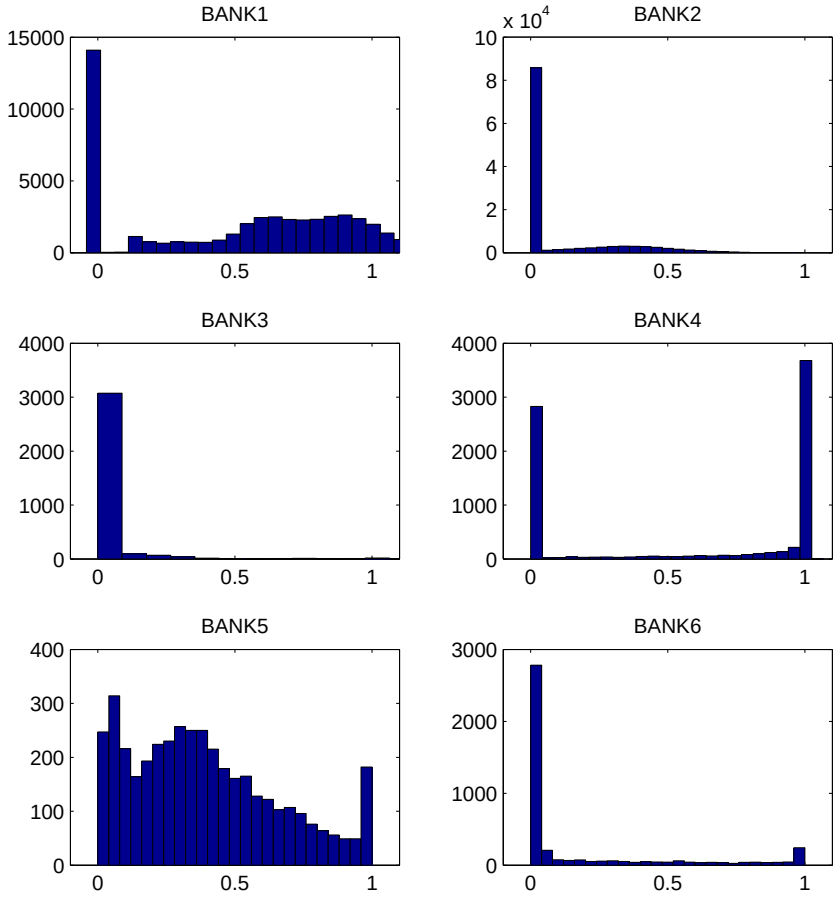

Figure 2.1: Empirical LGD distributions for six real-life datasets

diction performance of these models. The independent continuous variables are standardized with the sample mean and standard deviation of the training set. Further, independent nominal variables are transformed by introducing as many dummy variables as there are nominal categories. A dummy variable takes the value 1 or 0 to indicate the presence or absence of a specific nominal category. Finally, independent ordinal variables are transformed by introducing as many thermo variables as there are ordinal categories. A thermo variable takes the value 1 when a specific ordinal category or higher order category is present and 0 otherwise.

2.5.2 Algorithm configurations

OLS, B-OLS and BR can be run without the need for any parameter tuning. For RiR, the ridge parameter is tuned by ten-fold cross validation on the training set. Values are varied from zero to one in steps of 0.01, and mean squared error is used as selection criterion. For RoR, the commonly used bisquare function is chosen as objective function and its parameter k is set to 4.685 times the standard deviation of the residual (115). The value of the power parameter for the BC-OLS models is varied over a chosen range, i.e. from minus three to three in 0.25 increments, and an optimal value is chosen based on a maximum likelihood criterion.

For the RT model, the training set is further split into a training and a validation subset. The validation set is used to select the criterion for evaluating candidate splitting rules (i.e. variance reduction

or ProbF), the depth of the tree and the threshold p-value for the ProbF criterion. All were selected based on the mean squared error on the validation set. To run MARS, we set the penalty for adding a Hinge function to 2.5 (116); the maximum interaction degree is varied from zero to five in steps of one and a setting is chosen based on mean squared error using ten-fold cross validation on the training set.

For LSSVM regression, the radial basis function (RBF) kernel is used because of its good overall performance for LSSVM classifiers (117). Its hyperparameters are again tuned using ten-fold cross validation on the training dataset. A grid search procedure evaluates a large space of possible hyperparameter combinations so as to find a combination that minimizes the mean squared error. The limits of the grid for the kernel and regularization parameter are set to $\left[0.5\sqrt{n},\, 500\sqrt{n}\right]$ and $\left[\frac{0.01}{m}, \frac{1000}{m}\right]$, where n, m denote the number of observations and variables, respectively (118). On a larger dataset, this search process can be computationally intensive. Therefore, a random sample of 4000 observations is chosen from the complete training set for the purpose of tuning the LSSVM hyperparameters before the final model is run on the full training set.

In order to train the ANNs, we again split each training set into a training and validation set. Validation-set mean squared error is then used to select the target layer activation function (logistic, linear, exponential, reciprocal, square, sine, cosine, tanh or arcTan) and determine the number of hidden neurons (a range of one to

twenty is considered). The hidden layer activation function is set to logistic.

Further, input selection methods are used to remove irrelevant or redundant independent variables from the datasets as this may improve the performance of the resulting regression models. More specifically, a stepwise selection procedure is applied in building the linear models, i.e. OLS, B-OLS, BR, BC-OLS, RiR and RoR. For computational efficiency reasons, an R^2-based filter method (119) is applied prior to building the LSSVM and ANN models. Both RT and MARS already perform variable selection implicitly so no additional input selection is required here.

2.5.3 Model evaluation

The performance of the resulting models is measured on the test set according to the eight performance metrics. On each dataset, the techniques are ranked from one (best) to twenty-four (worst) based on the resulting values for each of these metrics. Then, the average rank of each technique over all datasets is calculated for each metric. To further summarize the results, an overall average ranking of techniques over the datasets and over all metrics is also produced.

Model performance is statistically compared using Friedman's test (81) and the post-hoc multiple testing procedure of Hommel (82) as suggested in the literature (79, 80). Friedman's test is performed to test the null hypothesis that all regression techniques perform

alike based on their ranking for a chosen performance metric. We then use Hommel's method to compare each regression technique against the best performing one and report significant rank differences. All statistical tests are conducted at the 95% confidence level.

2.5.4 Implementation details

The majority of regression techniques are implemented through standard methods available in both Matlab (LOG, OLS, RT, RiR, RoR) and SAS (ANN, BC-OLS, BC-OLS, BR). External Matlab toolboxes are used for LSSVM (LS-SVMlab) and MARS (ARESLab). Variable selection is performed through the sequential feature selection method in Matlab and the R^2-based filter method in SAS. Further, standard methods are available in Matlab for calculating correlation coefficients such as Pearson's r, Spearman's ρ and Kendall's τ. For the statistical comparison using Friedman's test and the post-hoc multiple testing procedure of Hommel, a standalone Java application is used which is provided by Garcia and Herrera (80). All other code required for the experiments is developed by the author.

2.6 Results and discussion

Tables B.1 to B.6 contain the performance results obtained for all techniques on the six respective datasets. The best performing model according to each metric is underlined. The Friedman test

results for the respective performance measures all indicate that the observed differences in performance are extremely unlikely to be due to random chance (e.g. for R^2, the p-value is $1.06{\cdot}10^{08}$). Figure 2.2 displays a series of box plots for the observed distributions of performance values for the metrics AUC, R^2, r, ρ and τ. No box plots are however constructed for $RMSE$, MAE and AOC since these are dataset dependent and thus not comparable accross different datasets. Similar trends can be observed across the box plots. Note that differences in type of portfolio, number of observations and available independent variables are the likely causes of the observed variability of actual performance levels between the six different datasets.

Although all performance metrics listed above are useful measures in their own right, it is common to use the coefficient of determination R^2 to compare model performance across different datasets. As shown in Figure 2.2, the average R^2 of the models varies from about 4% to 43% which is in line with the reported results in previous studies (25, 65, 67, 68). In other words, the variance in LGD that can be explained by the independent variables is consistently below 50%, implying that most of the variance cannot be explained even with the best models. Note that although R^2 usually is a number on a scale of zero to 1, R^2 can yield negative values for non-OLS models when the model predictions are worse than always using the mean from the training set as prediction.

Table 2.2 shows the average ranking of techniques over the six

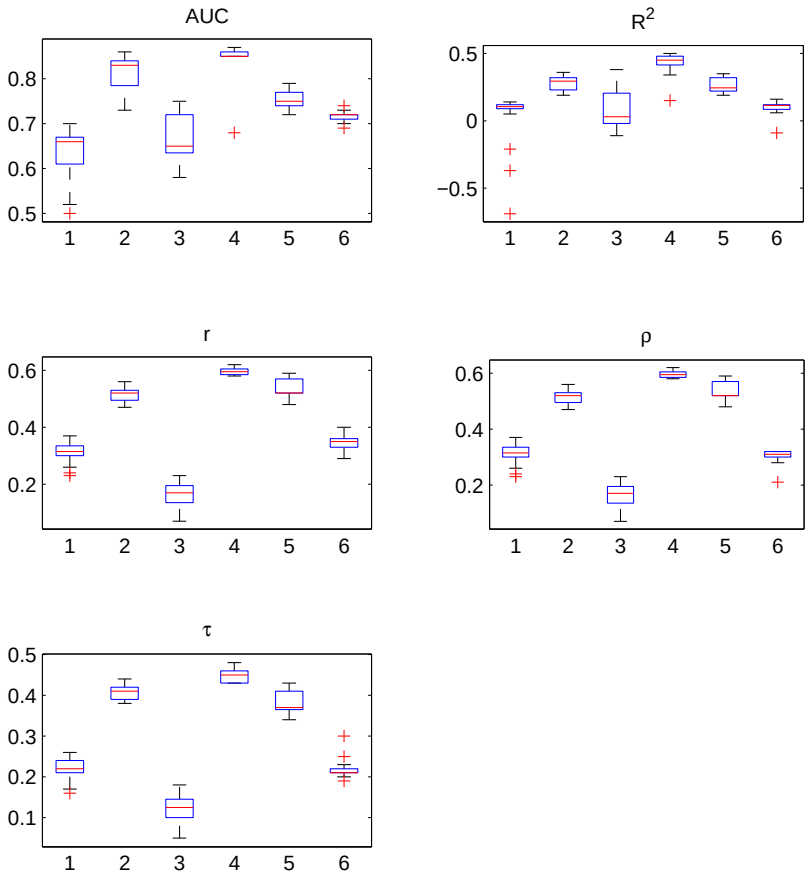

Figure 2.2: Variability of LGD model performance observed for the six datasets

Rank	Technique	RMSE	MAE	AUC	AOC	R^2	r	ρ	τ	OAR
1	LSSVM	4.0	9.0	3.8	4.0	4.5	4.2	3.5	4.8	4.7
2	ANN	3.5	3.7	6.3	3.2	3.8	3.8	8.8	8.9	5.3
3	OLS+LSSVM	5.3	9.6	4.1	5.0	5.8	5.8	5.5	6.2	5.9
4	LOG+ANN	5.1	7.0	10.0	5.4	5.7	5.2	7.8	8.3	6.8
5	OLS+ANN	5.7	9.3	4.6	7.8	6.8	5.8	7.7	7.8	6.9
6	LOG+LSSVM	8.2	8.4	5.9	8.3	8.8	6.2	6.7	6.6	7.4
7	OLS+MARS	5.4	13.4	5.5	5.1	5.9	5.3	10.0	11.0	7.7
8	OLS+RT	8.5	12.2	7.2	6.2	8.5	7.8	9.3	10.3	8.8
9	MARS	7.2	13.6	7.2	7.0	7.5	7.8	10.5	9.5	8.8
10	LOG+MARS	10.0	10.6	10.4	10.1	11.2	10.3	11.8	11.0	10.7
11	RT	10.1	11.4	*19.1*	10.4	10.5	11.2	10.8	6.5	11.2
12	LOG+RiR	13.2	14.7	12.6	13.0	15.1	14.2	11.0	11.8	13.2
13	LOG+RT	13.2	12.5	14.8	13.2	13.6	13.0	12.7	13.2	13.3
14	LOG+RoR	*17.3*	10.0	15.0	*17.3*	*18.2*	14.8	10.8	10.7	14.3
15	RiR	14.5	*20.2*	12.8	14.6	*15.8*	*15.8*	*16.8*	*17.4*	16.0
16	LOG+OLS	14.5	*16.8*	*16.3*	13.8	*15.7*	*16.5*	*16.8*	*17.5*	16.0
17	LOG+B-OLS	*17.6*	7.7	*17.3*	*17.7*	*18.5*	*16.7*	*18.6*	*19.3*	16.7
18	B-OLS	*20.4*	10.3	15.5	*21.2*	*15.8*	*20.2*	*16.0*	16.5	17.0
19	LOG+BC-OLS	*19.9*	10.3	*19.0*	*19.9*	*14.7*	*18.5*	*16.8*	*17.0*	17.0
20	OLS	*15.4*	*19.7*	13.8	*15.3*	*17.0*	*17.3*	*18.4*	*19.5*	17.1
21	RoR	*19.7*	*16.4*	*16.3*	*19.5*	*18.8*	*17.7*	*17.8*	14.0	17.5
22	BC-OLS	*22.1*	13.2	*20.7*	*22.1*	*16.7*	*20.8*	15.2	15.8	18.3
23	BR	*18.2*	*20.3*	*20.3*	*18.5*	*19.8*	*20.8*	*16.7*	16.7	18.9
24	LOG+BR	*21.3*	*19.8*	*21.6*	*21.3*	*21.4*	*20.5*	*20.0*	*19.7*	20.7

Table 2.2: Mean performance ranks of techniques over the six datasets and overall average rank (OAR) over all metrics

datasets according to each performance metric. Additionally, their overall average rank over the six datasets and over the eight performance metrics is included in the last column. The techniques are ordered according to their overall average ranking. The best performing technique for each metric is again underlined and techniques that perform significantly worse than this best technique according to Hommel's procedure are displayed in italic. It can be observed that the same techniques, LSSVM and ANN, are consistently ranked in the top two regardless of the metric.

The pure linear models built by OLS, RiR and RoR do not seem to

show consistent differences in performance between one another. Although RiR is ranked somewhat higher overall, it most often leads to model performance identical to that of OLS. We suspect that any potential benefits of RiR could be limited in this particular setting because the chosen variable selection methods eliminate highly correlated variables a priori. On all datasets, RoR produces models that either perform slightly worse than the OLS models or they show similar performance. Hence, RoR's ability to reduce the impact of outliers does not result in any actual performance improvement on our real-life datasets.

The linear models that incorporate some form of transformation to the dependent variable (i.e. B-OLS, BR, BC-OLS) are shown to perform consistently worse than OLS, despite the fact that these approaches are specifically designed to cope with the violation of the OLS normality assumption. This suggests that they too have difficulties dealing with the pronounced point densities observed in LGD datasets, while they may be less efficient than OLS or they could introduce model bias if a transformation is performed prior to OLS estimation (as is the case for B-OLS and BC-OLS).

Perhaps the most striking result is that, in contrast with prior benchmarking studies on classification models for PD (27), non-linear models such as LSSVM and ANN significantly outperform most linear models in the prediction of LGD. This implies that the relation between LGD and the independent variables in the datasets is non-linear (as is most apparent on dataset BANK3, see

Table B.3). Also, LSSVM and ANN generally perform better than RT or MARS. However, LSSVM and ANN result in black-box models while RT and MARS have the ability to produce comprehensible white-box models.

The performance evaluation of the class of two-stage models in which a logistic regression model is combined with a second-stage (linear or non-linear) model (LOG+), is less straightforward. Although a weak trend is noticeable that logistic regression combined with a linear model tends to increase the performance of the latter, it appears that logistic regression combined with a non-linear model slightly reduces the strong performance of the latter. Because the LGD distributions from BANK4, BANK5 and BANK6 also show a peak at $\text{LGD} = 1$, the performance of these models could possibly be increased by slightly altering the technique. Replacing the (binary) logistic regression component by an ordinal logistic regression model distinguishing between three classes ($\text{LGD} \leq 0$, $0 < \text{LGD} < 1$, $\text{LGD} \geq 1$) and then using a second-stage model for $0 < \text{LGD} < 1$ could perhaps better account for the presence of both peaks.

In contrast with the previous class of two-stage models, a clear trend can be observed for the combination of a linear and a non-linear model (OLS+). By estimating the error residual of an OLS model using a non-linear technique, the prediction performance tends to increase to somewhere very near the level of the corresponding one-stage non-linear technique. What makes these two-stage models attractive is that they have the advantage of combining the high

prediction performance of non-linear regression with the comprehensibility of a linear regression component.

2.7 Conclusions

In this chapter, twenty-four regression techniques were evaluated on six real-life datasets obtained from major international banking institutions. The average performance of the models in terms of R^2 ranged from 4% to 43%, showing that several resulting models have limited explanatory power. These rather weak performance results are quite similar to those obtained in previous LGD forecasting studies. Nonetheless, a clear trend can be seen that non-linear techniques, and support vector machines and artificial neural networks in particular, yield significantly higher model performance than more traditional linear techniques. This suggests the presence of non-linear relations between the independent variables and LGD, contrary to previous benchmarking studies on PD modeling where the difference between linear and non-linear techniques was not that explicit. Therefore, the study clearly demonstrated the potential of applying non-linear techniques to LGD modeling, possibly in the form of first order regression splines so as to yield good predictive performance while offering the advantage of being well interpretable.

3

Backtesting LGD models

"When I see articles with lots of significance tests,
I say that the statisticians are p-ing on the research."
-Herman Friedmann (American statistician, 1930-2010)

"The only relevant test of the validity of a hypothesis
is comparison of prediction with experience."
-Milton Friedman (American economist, 1912-2006)

The Basel accords require financial institutions to regularly validate their LGD models. This is crucial so banks are not underestimating or overestimating the minimal required capital to protect them against the risks they are facing through their lending policies. The validation of an LGD model typically includes backtesting which is the process of evaluating to which degree the internal model estimates correspond with the realized observations. Current backtesting practices are limited to solely measuring the similarity between model predictions and realized observations. It is however

not straightforward to determine acceptable performance based on these measurements. Although recent research lead to advanced backtesting methods for PD models, literature on similar backtesting methods for LGD models is non-existing. This study addresses this literature gap by proposing a backtesting framework with statistical hypothesis tests to support the validation of LGD models. The proposed statistical hypothesis tests implicitly define reliable reference values to determine acceptable performance and take into account the number of LGD observations which may influence the quality of the backtesting procedure. The workbench of statistical hypothesis tests is applied to an LGD model based on real-life data. Special attention is given to the evaluation of the statistical power of the proposed tests.

3.1 Introduction

Banks are required to regularly validate the internal estimation process and the internal models so as to prove their soundness to the national regulator (6). The validation of the estimation process involves issues like data quality, reporting and problem handling and how the predictive models are used by the bank. The validation of the estimation process is mainly qualitative in nature, although quantitative methods are useful for the examination of data quality. The validation of the models on the other hand includes both the examination of the model design and the predictions it produces. The evaluation of the model design consists of a qualitative review of the statistical techniques and the relevance of the data used to

build the model. The assessment of a model's predictions typically includes quantitative methods as benchmarking and backtesting.

While benchmarking methods evaluate the internal model estimates with external model estimates (88), backtesting methods evaluate the internal model estimates with the actual realized observations. The purpose of backtesting is to evaluate the predictive performance of a model and to assess its time evolution to detect model deterioration in a timely manner. An LGD model can experience reduced predictive performance when current loan loss behavior does not reflect previous loan loss behavior anymore on which the model is built. This may lead to an overestimation or underestimation of a bank's required minimal capital so that its operations can become less profitable or more risky respectively. Although banks are required to validate their models in order to be Basel compliant, the accord does not mention how to perform the validation (6). In addition, recent research has largely focused on advanced methods for backtesting PD models (70, 71, 72) but literature on comparable methods for backtesting LGD models is non-existing.

Current LGD backtesting practices are usually limited to comparing internal LGD predictions and realized LGD observations with error based metrics, correlation based metrics or even classification based metrics (88). It is however not straightforward to determine acceptable performance solely based on these metrics. A single value has little meaning without an appropriate reference value indicating acceptable accuracy. Additionally, these metrics do not take into ac-

count the number of LGD observations. When the portfolio lacks sufficient observations, a few extreme observations can distort the accuracy result and so degrade its reliability. This study proposes a backtesting framework where the model performance on test data is evaluated with respect to the model performance on training data with appropriate statistical hypothesis tests. Hence, an appropriate reference value is introduced while the number of observations is implicitly taken into account.

The remainder of this study is organized as follows. First, a literature review is conducted on empirical LGD studies which focus on the evaluation of the predictive performance of LGD models. Second, the key idea of the proposed backtesting procedure is explained together with the workbench of appropriate statistical hypothesis tests to evaluate LGD models. Third, the experimental set-up to apply and to evaluate the backtesting framework is described. This involves information about the employed real-life LGD data, the design of a predictive LGD model based on this data, a statistical significance analysis of the measured predictive model performance and a statistical power analysis of the proposed tests based on these performance metrics. Forth, the results of the backtesting procedure applied to a real-life LGD model is reported and discussed.

3.2 Literature review

The Basel accords require banks to backtest their internal models but do not further specify how this needs to be performed (6).

Current backtesting practices in the empirical LGD literature are usually limited to comparing internal LGD predictions and realized LGD observations with error based metrics (e.g. MAE, RMSE), correlation based metrics (e.g. Pearson's r, Kendall's τ, Spearman's ρ, coefficient of determination R^2) or even classification based metrics (e.g. AUROC) (88). Each of these metrics has its own method with respect to the way of quantifying the degree of similarity between LGD model predictions and the actual realized observations. This section describes the workings of these metrics more in detail and how these are used to assess the predictive performance of LGD models. To conclude, several problems are identified when using these metrics for the purpose of backtesting LGD.

Error based metrics quantify the error or difference between predicted and observed values. The most often used error based metric seems to be the MSE (25, 62, 64). The MSE is defined as the average of the squared difference between predictions and observations. Since errors are squared, this metric heavily weights outliers. The metric is bound between the maximum squared error and zero (perfect prediction). The RMSE is also often used as a metric in the literature (61, 63, 69). The RMSE is merely the squared root of the MSE but offers the additional advantage that it has the same units as the dependent variable being predicted, unlike MSE. Another error based metric used in the literature is the MAE (61, 63, 64). The MAE is given by the averaged absolute differences of predicted and observed values. Just like the RMSE, the MAE has the same unit scale as the dependent variable being predicted. Unlike RMSE,

MAE is not that sensitive to outliers. The metric is bound between the maximum absolute error and zero (perfect prediction).

Correlation based metrics quantify the degree of a statistical relationship between predicted and observed values. A very popular correlation based metric seems to be the R^2 (25, 65, 67, 68). The R^2 can be defined as one minus the fraction of the sum of squared errors to the variance of the observations. Since the second term in the formula can be seen as the fraction of unexplained variance, the R^2 can be interpreted as the fraction of explained variance. Although R^2 is usually expressed as a number on a scale from zero to one, R^2 can yield negative values when the model predictions are worse than using the mean $\overline{y}$ from the training set as prediction. Other correlation based metrics include Pearson's r (62), Spearman's ρ (88) and Kendall's τ (66). Pearson's r measures the degree of linear relationship between predictions and observations. Spearman's ρ is defined as Pearson's r applied to the rankings of predicted and observed values. Likewise, Kendall's τ measures a similar degree of correspondence of the ranked ordenings between predictions and observations. All three correlation coefficients can take values between minus one (perfect negative correlation) and one (perfect positive correlation) with zero meaning no correlation at all.

Although not considered to be a metric to assess the performance of a regression model, a typical binary classification based metric as the Area Under the Receiver Operating Characteristic curve (AUROC) (89) is used in the LGD literature (62, 66, 68). It is employed

in an LGD context to measure how good an LGD regression model is able to distinguish between high and low losses. To build the ROC curve, the observed values are first classified into high and low classes using the mean $\overline{y}$ of the training set as reference. The area under the ROC curve is an estimate for the discriminatory power of a model. The metric varies from 0.5 (random classification) to one (perfect classification). Another similar metric is the Area Over the Regression Error Characteristic curve (AOREC) (86). It can be seen as either a generalization of an error based metric or a generalization of the AUROC. The AOC curve plots the error tolerance on the x-axis versus the percentage of points predicted within the tolerance (or accuracy) on the y-axis. The resulting curve estimates the cumulative distribution function of the squared error. The area over the REC curve (AOC) is an estimate of the predictive power of the technique. The metric is bound between zero (perfect prediction) and the maximum squared error.

The evaluation schema to assess the predictive performance of an LGD models varies in the literature. For prediction it is important that the model performance is evaluated on unseen cases which it will also encounter in real-life. These evaluation schema's are called out-of-sample. In an out-of-sample schema (61, 63, 65, 66, 68), the LGD dataset is split in a random training set (typically two-third of the total dataset) and a test set (remaining one-third of the total dataset). The training set is used to build the model and the test set is used to evaluate the model. In order to enhance the reliability of the assessment, multiple hold-out validations are often

executed (61, 63). A more strict out-of-sample evaluation schema is also out-of-time. In an out-of-time schema (25, 61, 62, 64, 69), the model is built on data of a specific time period and is evaluated on data after this time period. While an average of multiple hold-out validations is most applicable to assess how good a technique fits a model to a dataset, an out-of-time validation is most applicable to assess the real-life predictive model performance as the model is strictly built using historical data and strictly evaluated on future data. Backtesting always comes down to an out-of-time evaluation.

The use of the above described metrics for backtesting an LGD model may cause flaws. First of all, it is not straightforward to determine acceptable model performance solely based on these metrics. A single value has little meaning without an appropriate reference value indicating acceptable performance. For example, an LGD model performance of 50% in terms of R^2 may sound bad since a perfect LGD model should correspond with an R^2 of 100%. However, comparing this performance with other real-life LGD benchmarking results where the average R^2 ranges from 4% to 43% (88), this may sound very good. In addition, these metrics do not take into account the number of LGD observations. When the portfolio lacks sufficient observations, a small amount of extreme observations can distort the accuracy results and so degrade its reliability. For example, when assessing an LGD model performance in a specific year containing only ten defaults in the portfolio, a few extreme bad model predictions may cause a disproporationate low performance.

3.3 Proposed backtesting framework

The proposed key idea to backtest the predictive performance of an LGD model is to evaluate the model performance on the test data with respect to the model performance on the training data with appropriate statistical hypothesis tests. By comparing the model test performance with the model training performance, a reference value is introduced, tailored to the respective model. Model deterioration is thus defined as a decrease of model performance compared to the performance during model building. Note that this is in contrast to the process of benchmarking where the performance of multiple models is compared with each other. By applying statistical hypothesis tests, model deterioration can be statistically detected with a pre-defined significance level (e.g. de facto 5%). In addition, statistical hypothesis tests implicitly take into account any insufficient number of observations (i.e. sample size) to prevent incorrect judgements.

In what follows, the proposed statistical hypothesis tests to decide upon acceptable model performance are explained. These tests typically start with the formulation of a null hypothesis H_0 which assumes no model deterioration and an alternative hypothesis H_a which indicates model deterioration. Further, a test statistic T is identified in order to assess the truth of H_0. A decision whether or not to reject H_0 can be made by calculating the test statistic T on the concerning sample and to compare this to the critical value corresponding to a significance level of 5%. If the resulting test statistic is at least as extreme than the critical value, H_0 may be

rejected in favor of H_a, otherwise H_0 may not be assumed.

3.3.1 Central tendency error tests

The most basic model performance metric is the central tendency of the error. This is sometimes referred to as model calibration. The error E is defined as the difference between predictions $\hat{Y}$ and observations Y or $E = \hat{Y} - Y$. Two well-known statistical hypothesis tests in the literature may be used for this purpose: the T test and the Wilcoxon signed rank test. Both tests allow to evaluate to what degree the central tendency of the error equals zero which serves a the reference value. It is assumed that the central tendency of the training error of a well-aligned model equals zero. While the T test compares the mean error to zero, the Wilcoxon signed rank test compares the median error to zero. Note that one-tailed tests are used instead of two-tailed tests because these provide more power to detect whether the average prediction is lower than the average observation by not testing the opposite. An underestimation of average loss may be fatal for a bank but an overestimation may merely increase its capital requirements.

The T test determines to what degree the mean of the error μ_E equals zero:

$$H_0 : \mu_E = 0,\ H_a : \mu_E < 0$$

The test statistic T can be derived from the Central Limit Theorem which states that the sample mean $\overline{e}$ converges to a normal

distribution and Cochran's theorem which states that the sample deviation s_e is χ^2_{n-1}-distributed. Hence, in that case the resulting test statistic follows a t_{n-1}-distribution:

$$T = \frac{\bar{e}}{\frac{s_e}{\sqrt{n}}} \sim \frac{Z}{\sqrt{\frac{\chi^2_{n-1}}{n-1}}} \sim t_{n-1}$$

with n the number of predictions to backtest. Note that when n is large (i.e. $n > 30$), a χ^2_{n-1}-distribution converges to a normal distribution. Hence, the resulting test statistic also follows a normal distribution and performing a Z test is equally appropriate.

The Wilcoxon signed rank test (85) determines to what degree the median of the error equals zero:

$$H_0 : \eta_E = 0, \;\; H_a : \eta_E < 0$$

The test statistic T can be derived by calculating the sum of the positive ranked errors r_+. The positive ranked errors are determined as follows. Zero errors are ignored, the smallest positive error is ranked 1, the next smallest positive error is ranked 2, etc. In case of ties, average ranks are assigned. The resulting test statistic approximates a normal distribution according to the Lyapunov Central Limit Theorem:

$$T = \frac{r_+ - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \sim N(0,1)$$

Compared to the T test which draws conclusions based on the abso-

lute value of the mean of the test sample, the Wilcoxon test statistic implicitly determines the deviation of the central tendency of the error from zero. Nonetheless, in order to be able to easily quantify and compare the central tendency error over the test years, the Wilcoxon metric w_r is defined and used in what follows. This is the ratio of the sum of positive ranked errors (r_+) to the total sum of both negative and positive ranked errors ($r_+ + r-$). It is bound between zero (underestimation) and one (overestimation) with 0.5 indicating zero central tendency error.

3.3.2 Dispersion error tests

Next to the central tendency of the error, a complementary basic model performance metric is the dispersion of the error. This is sometimes referred to as model precision. Two well-known statistical hypothesis tests may be used for this purpose: the F test and the Ansari-Bradley test. Both tests allow to evaluate to what degree the dispersion of the error differs from the dispersion of the training error which serves as a reference. While the F test compares the variance of the error with the variance of the training error, the Ansari-Bradley test measures and compares the dispersions of the error and training errors by leaning upon rankings rather than on the numeric values of the data. Note that also here one-tailed tests are proposed to enhance the statistical power to detect when the dispersion of the error is larger than the dispersion of the training error. A larger dispersion may cause more unforeseen losses.

The F test (83) determines to what degree the variance of the error

σ_E^2 is equal to the variance of the training error $\sigma_{E_t}^2$:

$$H_0 : \sigma_E^2 = \sigma_{E_t}^2,\ H_a :\ \sigma_E^2 > \sigma_{E_t}^2$$

The test statistic T can be derived by inspecting the ratio of the variances. According to Cochran's theorem, the sample variances s_e^2 and $s_{e_t}^2$ follow a χ^2-distribution with $n-1$ and n_t-1 degrees of freedom, respectively. Hence, the resulting test statistic follows an F-distribution with $n-1$ and n_t-1 degrees of freedom:

$$T = \frac{s_e^2}{s_{e_t}^2} \sim \frac{\left(\frac{\chi_{n-1}^2}{n-1}\right)}{\left(\frac{\chi_{n_t-1}^2}{n_t-1}\right)} \sim F_{n-1,n_t-1}$$

with n the number of defaults to backtest and n_t the number of defaults to train the model.

The Ansari-Bradley test (84) determines to what degree the cumulative distribution function of the error $F_E(u)$ and the cumulative distribution function of the training errors $F_{E_t}(u)$ are equal, assuming they can only differ in the value of a scale parameter θ:

$$H_0 : F_E(u) = F_{E_t}(u),\ H_a : F_E(\theta u) = F_{E_t}(u)\ with\ \theta > 1$$

The test statistic T can be derived by calculating the sum of weights of the ordered errors of the combined sample e and e_t with total size $m = n + n_t$. The weights assigned are one to both the smallest and largest error in the combined sample, 2 to the next smallest and next largest, etc., $\frac{m}{2}$ to the two middle observations if m is even, and $\frac{m+1}{2}$ to the one middle observation if m is odd. The resulting test

statistic is the sum of weights of the ordered errors in the combined sample associated with e, defined as w_e, and approximates a normal distribution according to Ansari and Bradley:

$$T = \frac{w_e - \frac{n(m+2)}{4}}{\sqrt{\frac{nn_t(m+2)(m-2)}{48(m-1)}}} \sim N(0,1)$$

when m is even, or:

$$T = \frac{w_e - \frac{n(m+1)^2}{4m}}{\sqrt{\frac{nn_t(m+1)(3+m^2)}{48m^2}}} \sim N(0,1)$$

when m is odd. Although the test requires that E and E_t have identical population medians, Ansari and Bradley recommend subtracting the sample medians and shift both e and e_t to zero median if this assumption should not be met.

Compared to the F test which draws conclusions based on absolute values of training and test sample variances, the Ansari-Bradley test statistic implicitly determines the gap between training and test sample dispersion. Nonetheless, in order to be able to easily quantify and compare the test performances over the test years albeit relative to the training performance, the Ansari-Bradley metric ab_w is defined and used in what follows. This is the ratio of the sum of weights of the ordered errors in the combined sample associated with e (w_e) to the total sum of weights of the ordered ranks in the combined sample associated with both e and e_t ($w_e + w_{e_t}$). It is

bound between zero (larger dispersion) and one (lower dispersion) with 0.5 indicating similar dispersion error.

3.3.3 Error, correlation and classification based tests

In addition to the central tendency and dispersion of the error, other metrics are frequently used in the empirical LGD literature to assess model performance. These are error based (i.e. $RMSE$, MAE, $AOREC$), correlation based (i.e. R^2, r, ρ, τ) or classification based (i.e. $AUROC$) metrics. However, there are no statistical hypothesis tests described in the literature on how these may be used to detect model deterioration. The main problem is that it is not straightforward to determine the theoretic distribution of a test statistic under a null hypothesis based on these metrics. Nonetheless, such a distribution may be estimated via a bootstrapping approach. The basic idea of bootstrapping is that inference of a population from sample data can be modeled by inference sample data from resampling the sample data. For this purpose, it allows to empirically construct a distribution of a test statistic under a null hypothesis when this is theoretically unknown.

A bootstrap test determines to what degree the performance P is equal to the training performance P_t:

$$H_0 : P = P_t,\ H_a : P > P_t$$

where the test statistic T is defined as $P_t - P$ and where P may be one of the commonly used LGD model performance metrics listed

above. The distribution of the test statistic under the null hypothesis can be simulated through bootstrapping according to the Beran algorithm (120, 121, 122, 123). First, the training and test observations are stacked together as well as the training and test predictions. Next, a training/test bootstrap sample with the same length as the original training/test set is extracted from the stacked observations/predictions through random sampling with replacement. Then, the difference of the concerning metric for the bootstrap training sample and bootstrap test sample is calculated. This procedure is repeated (e.g. de facto about 1000 times) in order to empirically build up the distribution of the test statistic under the null hypothesis. Note again that only one-tailed tests are proposed to enhance the statistical power to detect performance deterioration.

3.4 Methods

This section describes the evaluation of the proposed backtesting framework applied on a real-life LGD model. The experimental set-up is as follows. First, real-life loss data is collected consisting of a variety of characteristics of the respective loans on the one hand and the corresponding observed LGD on the other hand. Second, a regression analysis of the loss data is performed in order to construct a predictive LGD model. Third, the performance of the predictive LGD model is out-of-time backtested on multiple years. For this purpose the proposed statistical hypothesis tests are performed in order to discover any significant model deteriorations. Forth, the proposed statistical hypothesis tests are empirically evaluated

through a statistical power analysis.

3.4.1 Data collection

The real-life LGD dataset collected in this study reflects corporate loan loss over a time span from 1984 to 2004 and contains 891 observations. Data from 2001 to 2004 is used to yearly backtest the constructed LGD model. The model is built with data from 1984 to 2000. This split between training and test data on 2000 is chosen so as to have sufficient data (e.g. about 500 defaults) to train an LGD model while still having sufficient time periods (i.e. four years) to backtest the LGD model. The number of observations used for training and backtesting purposes is given in Table 3.1.

Year	Observations	Purpose
2004	30	Backtesting
2003	47	
2002	140	
2001	155	
1984-2000	519	Training

Table 3.1: Number of observations

The distribution of the LGD data used for both training and testing is illustrated in Figure 3.1. This appears to be predominantly J-shaped with the highest frequency at the end of the range. This means that the dataset is characterized by high LGDs caused by the majority of the defaults. Notice that especially 2001 and 2002 are characterized with high LGDs while this shifts to generally lower LGDs for 2003 and 2004. Based on the literature, the LGD distri-

bution is indeed typically non-normal distributed and most often rather bimodal distributed. Real-life LGD tends to be characterized by high concentrations of either total recovery or total loss or both. The majority of the empirical LGD literature reports of a large peak on zero and a smaller peak on one (61, 62, 64, 66, 67). Nonetheless, few studies also report the opposite as is also the case for this dataset: a large peak on one and a smaller or non-existing peak on zero (62, 65).

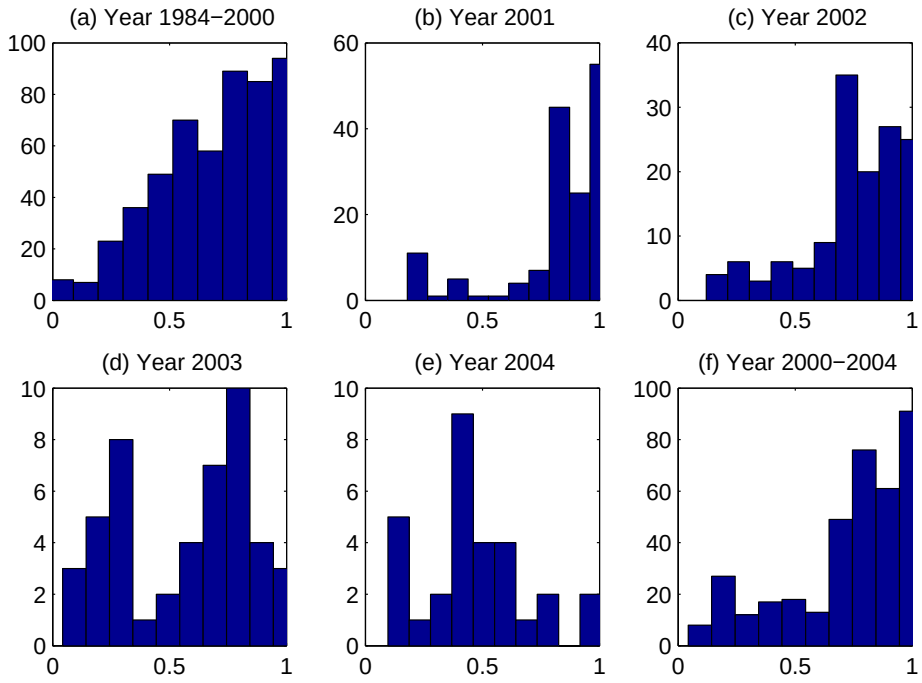

Figure 3.1: LGD observations histogram

The LGD dataset covers both loans and bonds from large corporates in the USA. Next to the LGD target variable, the dataset includes 42 variables which represent potential LGD drivers, a.o.

rating, level of seniority, country of domicile, type of industry, default rate. The data covers different sectors such as transportation, finance, public, industrial and real estate. Domiciles are located in America, Europe and Oceania. For the purpose of predictive modeling, a few pre-processing actions are executed. Continuous variables are transformed to the standard z-score with the sample mean and standard deviation of the training set. Furthermore, categorical variables are quantified by dummy encoding. More information about this dataset is confidential.

3.4.2 Predictive modeling

In first instance, a predictive LGD model is required to estimate future outcomes as well as possible. This allows banks to protect themselves against default risks and to remain competitive. In second instance, banks need to provide comprehensible LGD models. This is required by the national regulators in order to ensure that banks fully understand their risks and underlying model relations. Although non-linear models such as Support Vector Machines and Artificial Neural Networks seem to show significantly higher performance on average than linear models, these are labeled as being non-comprehensible (88). National regulators may not allow hard to interpret models since financial institutions may be legally obliged to motivate why a customer is denied credit (124). Therefore, for our research purposes, it is deliberately chosen to deploy a simple linear model to obtain the most understandable model form in order to fully interpret its backtesting results. Note that, for the

purpose of this backtesting exercise, the model performance in absolute terms actually does not play any role.

Based on the real-life LGD dataset, a linear model is determined by minimizing the sum of squared differences between predictions and observations of the training set. In order to increase the generalization behavior, i.e. the ability to estimate the LGD on out-of-sample data, a variable selection method is used to exclude irrelevant or redundant variables from the model. Based on a ten fold holdout validation schema, a model wrapper searches for a subset of variables that best predicts the LGD by sequentially selecting variables until there is no improvement in minimizing the sum of squared differences between predictions and observations. The selected subset includes two binary variables referring to the level of seniority, i.e. senior unsecured (SU) and junior subordinated (JS), and one continuous variable, i.e. US default rate from the previous year (USDR(t-1)). The output of the variable selection strengthens previous literature studies which stress the importance of seniority and default rate as major predictive drivers (6):

$$\mathbf{LGD} = 0.74 - 0.15 \cdot \mathbf{SU} + 0.18 \cdot \mathbf{JS} + 0.02 \cdot \mathbf{USDR(t\text{-}1)}$$

The resulting linear model can be interpreted as follows. The baseline LGD is 74% and decreases with 15% when the loan is senior unsecured or increases with 18% when the loan is junior subordinated. Additionally, the LGD increases with the US default rate from the previous year with a speed of 2% per unit. These relations are in line with previous empirical studies. Secured debt and high priority de-

crease the LGD (45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55). Note that the seniority of the loan dominates over the security in this model since the SU dummy decreases the LGD. Further, LGD is typically higher in a period of high defaults (47, 48, 49, 54, 55, 56, 57).

3.4.3 Significance analysis

Table 3.2 gives an overview of the performance metrics on which the statistical hypothesis tests of the proposed backtesting framework are based. The first two metrics specifically measure the central tendency of the error while the subsequent two metrics specifically measure the dispersion of the error. Standard (non-)parametric tests are available in literature to test performance deterioration in terms of these metrics. The following eight metrics are quite diverse and have their own specific method of quantifying the degree of similarity between predictions and observations. No standard tests are however available in literature to detect performance deterioration based on these metrics. Nonetheless, the proposed bootstrap based tests can offer relief here.

The minimal and maximal performance values of the corresponding metrics are given in columns two and three of Table 3.2. Although R^2 can yield excessive negative values when the model predictions are worse than using the mean from the training set as prediction, these have however the same meaning as zero values, i.e. that the model does not explain any variation at all (125). Hence, any negative values are replaced by zero to enhance its interpretation and to prevent distortion of the corresponding bootstrap tests. Note that

strictly seen, $AUROC$ can also yield values between 0 and 0.5 and r, ρ and τ can also take negative values from -1 to 0. This may occur in case of negative classification performance or negative correlations, respectively.

Metric	Worst	Best
$\overline{e}$	-∞	0
w_r	0	0.5
s_e^2	$+\infty$	0
ab_w	0	0.5
RMSE	$+\infty$	0
MAE	$+\infty$	0
AUROC	0.5	1
AOREC	$+\infty$	0
R^2	0	1
r	0	1
ρ	0	1
τ	0	1

Table 3.2: Performance metrics

In order to decide upon acceptable performance for the metrics described above, the out-of-time performance is compared with the training performance. Each statistical hypothesis test assumes a null hypothesis and if sufficient evidence exists against the null hypothesis, the alternative hypothesis is concluded. This evidence is gathered in the form of a p-value. The p-value is the probability of obtaining a test statistic at least as extreme as the one that was actually observed, assuming that the null hypothesis is true. When the resulting p-value is compared to a pre-defined significance level, a decision can be made on statistical significance. The pre-defined

significance level is the probability of making a type I error (i.e. the incorrect rejection of the null hypothesis). This is generally denoted as α and de facto pre-defined as 5% (126). Low p-values (i.e. <5%) indicate that H_0 can be more confidently rejected, whereas high p-values (i.e. >5%) indicate that there is insufficient evidence to do so.

Note that a significance analysis may be extended in various ways. First, statistical comparisons may also be performed between the performance of the concerning test year and the performance of any previous year(s) instead of the performance on the training set, if required. Second, the statistical tests may also be performed on specific segments of the data. This segmentation could be either done on the input data (e.g. different levels of seniority or security) or on the output data (i.e. different levels from low to high LGD). Third, a traffic lights approach may be used to support the visualization of the resulting p-values. Different colors can be assigned to a specific range of p-values (70). The choice of the different ranges of p-values can however be decided by the financial institution. In addition, the kind and number of colors can also be chosen at the discretion of the financial institution, although a minimum satisfactory number of three is suggested (127). These extensions are however not put into practice for this study for reasons of clarity.

3.4.4 Power analysis

In order to evaluate whether the results of the statistical hypothesis tests are sufficiently reliable, the statistical power π is empirically

determined. The power of a test is defined as the probability that the test rejects the null hypothesis when this is indeed false. Note that this is the probability of not making a type II error (i.e. the failure to reject the null hypothesis while it is actually false). The probability of making a type II error is generally denoted as β. To decide upon acceptable statistical power, a de facto threshold of 85% is used (126). A test is considered to be sufficiently powerful when π is higher than 85% or β is lower than 15%. Note that β and thus also π is related to α. When α is higher, β is lower or π is higher, and vice versa.

The statistical power of a test is determined according to the Beran algorithm (120, 121, 122, 123). First, the alternative hypothesis distribution is empirically built. Therefore, a training/test bootstrap sample is extracted from the original training/test set with the same size through random sampling with replacement. Subsequently, the test statistic T is calculated on the bootstrap samples. This procedure is repeated about 1000 times as a rule of thumb in order to empirically build up a reliable distribution of the test statistic under the alternative hypothesis. Second, the probability of making a type II error β is calculated. Therefore, the critical value corresponding with the 95th (i.e. $1-\alpha$) percentile of the null distribution is determined. Then, the difference between the percentile of the alternative distribution corresponding with this critical value and the 0th percentile of the alternative distribution equals to β. Finally, the power can be calculated as $\pi = 1 - \beta$.

3.4.5 Implementation details

The standard parametric and non-parametric statistical hypothesis tests are implemented through standard methods in Matlab (T test, Wilcoxon test, F test, Ansari-Bradley test). The linear regression method and the sequential feature selection method in Matlab are used for model building and variable selection respectively. In addition, standard methods in Matlab are also used for calculating correlation coefficients as Pearson's r, Spearman's ρ and Kendall's τ. All other code required for the experiments is developed by the author.

3.5 Results and discussion

Metric	1984-2000	2001	2002	2003	2004
$\overline{e}$	0.00	-0.17	-0.12	0.08	0.16
w_r	0.43	0.15	0.20	0.53	0.83
s_e^2	0.05	0.05	0.05	0.07	0.05
ab_w	0.50	0.24	0.21	0.06	0.05
RMSE	0.23	0.29	0.25	0.28	0.27
MAE	0.18	0.26	0.22	0.25	0.23
AUROC	0.70	0.56	0.55	0.63	0.55
AOREC	0.05	0.08	0.06	0.08	0.07
R^2	0.12	0.00	0.00	0.01	0.00
r	0.34	0.14	0.19	0.30	0.17
ρ	0.33	0.03	0.22	0.24	0.07
τ	0.23	0.03	0.18	0.19	0.06

Table 3.3: Performance values

Test	2001	2002	2003	2004
T	**0.00**	**0.00**	0.97	1.00
W	**0.00**	**0.00**	0.98	1.00
F	0.43	0.73	**0.04**	0.52
AB	0.86	0.29	**0.00**	0.10
RMSE	**0.00**	0.09	**0.03**	0.07
MAE	**0.00**	**0.00**	**0.00**	0.06
AUROC	**0.00**	**0.00**	0.21	0.10
AOREC	**0.00**	0.07	**0.02**	0.07
R^2	**0.00**	**0.00**	0.13	0.18
r	**0.01**	0.05	0.33	0.15
ρ	**0.00**	0.13	0.25	0.10
τ	**0.00**	0.26	0.35	0.08

Table 3.4: Statistical significance values

This section reports and discusses the performance values of the LGD model, the statistical significance values of the performance differences between training and test sets and the statistical power values of the applied statistical hypothesis tests. The performance results of the LGD model for each metric are represented in Table 3.3. Both training (i.e. data from 1984 to 2000) and test set performances (i.e. data from 2001 to 2004) are given in order to see the evolution of the performances of the subsequent years with respect to the training performance. In order to detect significant performance deteriorations based on these performance values, Table 3.4 represents the resulting p-values of the appropriate statistical hypothesis tests corresponding to each performance metric. Finally, table 3.5 lists the power values of each statistical hypothesis tests so as to evaluate to what degree these are sufficiently reliable to discover performance deteriorations.

Test	2001	2002	2003	2004
T	**1.00**	**1.00**	0.00	0.00
W	**1.00**	**1.00**	0.00	0.00
F	0.09	0.02	0.54	0.01
AB	0.05	0.15	**0.94**	0.30
RMSE	**1.00**	0.36	0.79	0.40
MAE	**1.00**	**0.95**	**0.92**	0.57
AUROC	**0.87**	**0.95**	0.15	0.19
AOREC	**1.00**	0.39	0.75	0.39
R^2	**0.98**	**0.91**	0.13	0.09
r	**0.89**	0.50	0.06	0.38
ρ	**0.96**	0.27	0.15	0.50
τ	**0.95**	0.18	0.12	0.55

Table 3.5: Statistical power values

The evolution of the central tendency of the error in terms of the mean error $\overline{e}$ or the Wilcoxon metric w_r is represented in the first and second row of Table 3.3. Regardless of measuring the central tendency of the error with $\overline{e}$ or w_r, the same trend is extracted. The central tendency is below zero in terms of $\overline{e}$ and below 0.5 in terms of w_r for 2001 and 2002 while it is above zero in terms of $\overline{e}$ and above 0.5 in terms of w_r for 2003 and 2004. The corresponding p-values in Table 3.4 for both the T test and one sample Wilcoxon test equal to zero for 2001 and 2002 and are (close to) one for 2003 and 2004. This means that both tests agree that the model is significantly underestimating LGD for 2001 and 2002 while this is not the case for 2003 and 2004. The consistent underestimations of the model may point to more severe economic downturn period than expected. The corresponding power values in Table 3.5 for both the T test and one

sample Wilcoxon test are maximal for 2001 and 2002 and minimal for 2003 and 2004. This means that the detection of significant underestimations of LGD is supported by the large power in 2001 and 2002 and that the model is overestimating LGDs for 2003 and 2004. Notice that the Wilcoxon metric w_r from the training errors does not equal to 0.5 which ideally should be when there is zero central tendency of the error. This small gap is because the error is non-normally distributed which leads to a difference between the median error and the mean error which actually is equal to zero.

The evolution of the dispersion of the error in terms of the variance of the error s_e^2 or the Ansari-Bradley metric ab_w is shown in the third and forth row of Table 3.3. According to s_e^2, the dispersion of the error remains rather constant for the subsequent years, except for 2003 which shows an increased dispersion of the error. According to ab_w on the other hand, the dispersion of the error slightly degrades. The corresponding p-values in Table 3.4 for both the F test and Ansari-Bradley test are above the significance level of 5% except for 2003. This means both tests agree that there is only a significant deterioration of the dispersion error for 2003. The corresponding power values in Table 3.5 are low for both the F test and Ansari-Bradley test except for 2003. These low values however undermine the p-values pointing out no significant differences. This means we can not conclude with much certainty that there is no deterioration of the dispersion error. Nonetheless, the detection of significant differences shown for 2003 is supported by increased power of both tests for that year.

The evolution of the metrics which impersonate either a degree of error, classification or correlation are shown in the last eight rows of Table 3.3. For 2001, it can be seen that the test performance for 2001 is smaller than the training performance according to all these metrics. The corresponding p-values for 2001 in Table 3.4 are also all below the significance level of 5%. This means that all tests unanimously agree that there is a significant deterioration of the performance for 2001. For 2002 and 2003 however, some metrics still agree on significant performance deterioration although there is no unanimity. For 2004, no significant performance deteriorations could be detected although all metrics show consistently lower test performance with respect to the training performance. The corresponding power values in Table 3.5 are generally high for 2001 and slightly decrease for the subsequent years. The detected significant differences for the bootstrap tests are backed up by large power values. However, in the rest of the cases the bootstrap tests show moderate power when no significant differences are detected. This leaves decisions about performance deterioration in those years rather unconclusive.

When taking into account the resulting performance values, the p-values and power values, one can conclude that the model shows significant weak performance in 2001 but slightly shows improved performance during 2002 and 2003 to show rather good performance in 2004. The model performance deterioration behavior may be linked with the high number of defaults for 2001 which decrease

for the subsequent years as can be observed in Table 3.1. Higher default rates can lead to higher LGDs (128). This relation may be strengthened when observing the histograms in Figure 3.1 where a shift is noticed from high LGDs and a large number of defaults in 2001 and 2002 to lower LGDs and a small number of defaults. The higher default rates in 2001 and 2002 may be ascribed to the big recession period in the USA around the late 2000s, although the model takes into account the US default rate as macro-economic factor and is trained with data during a previous USA recession of the 1990s. The subsequent recovery period may explain the slow performance correction for these years. Generally, when the model is well trained and deteriorates over time, it means that the original training data is no longer representative for the current population. This can be caused by external changes (e.g. new developments in the economic, political or legal environment) or internal changes (e.g. new business strategies, exploration of new market segments or new organizational structure) (70). A data stability analysis may offer more insight into which variables cause possible shifts (70). In this case it is advised to build a new model with more representative training data.

3.6 Conclusions

This study addresses the call for more research on backtesting LGD models, a Basel validation requirement for any bank implementing the advanced IRB approach. Current backtesting practices often consist of measuring the similarity between model predictions

and realized observations. It is however not straightforward to determine upon acceptable model performance solely based on these metrics. First, a single value has little meaning without an appropriate reference value indicating acceptable accuracy. Second, when the portfolio lacks sufficient observations, a few extreme observations can distort the performance results and so degrade its reliability. This study proposes a framework to backtest LGD test with statistical hypothesis tests. The key idea is to evaluate the model performance on the test data with respect to the model performance on the training data with appropriate statistical hypothesis tests. Hence, an appropriate reference values is introduced while the number of observations is implicitly taken into account. For professionals, it is advised to backtest LGD models in three steps. First, the model performance needs to be measured with metrics of choice in order to see its evolution over the years. Second, corresponding statistical tests need to be performed to check for any significant deteriorations. Third, the power of each test needs to be quantified in order to assure if the test is sufficiently reliable when no significant deterioration is detected. The proposed backtesting framework is illustrated by backtesting an LGD model based on real-life loss rate rate data.

4

Selecting LGD models

"All models are wrong, but some are useful."
-George Box (British statistician, 1919-today)

"The proof of the pudding is in the eating.
By a small sample, we may judge of the whole piece."
-Miguel de Cervantes (Spanish novelist, 1547-1616)

Although techniques such as Support Vector Machines and Artificial Neural Networks show superior accuracy on 6 real-life LGD datasets, these are as such not suited for real-life LGD modeling because of their lack of comprehensibility which is a key requirement. This chapter presents a set of techniques which produce humanly interpretable models, i.e. linear, spline, tree, linear tree and spline tree, which can be used for real-life LGD modeling. Unfortunately, no model form is superior for all kind of regression datasets in general and LGD datasets in particular. Some studies claim that some

regression techniques are better suited for LGD modeling given its typical non-normal distribution characteristics. Apart from LGD research, other studies claim that also other typical dataset characteristics may favor a specific model form. Nonetheless, sufficient evidence remains absent. In this large-scale meta-learning study is explored in what degree dataset characteristics can predict which comprehensible model will fit a given dataset best. Since the very limited number of publicly available datasets, let alone LGD datasets, the experiments are conducted with more than thousand so called datasetoids representing various real-life dependencies to discover possible relations. It is found that algorithm based characteristics such as sampling landmarks are major drivers for successfully predicting the most accurate algorithm. Further, it is ascertained that data based characteristics such as the length, dimensionality and composition of the independent variables, or the asymmetry and dispersion of the dependent variable do not matter for this purpose.

4.1 Introduction

According to the benchmarking study in Chapter 2 involving six real-life LGD datasets, black box models built by Support Vector Machines provide significantly better fits on average than white box models such as for example built by Ordinary Least Squares. However, black box techniques as such are not suited for LGD modeling because of their lack of comprehensibility. National regulators may not allow hard to interpret models since financial institutions may be legally obliged to motivate why a customer is denied credit (124). Note that, next to domain of credit risk, it is often of crucial im-

portance both to obtain correct outcomes and to understand how a model comes to its conclusions. For example, in medical diagnosis it is important to gain insight in how certain variables may have an impact on the degree of a disease as these may provide valuable information about a potential cure (129).

Based on the techniques employed in Chapter 2, a selection of techniques to built humanly interpretable LGD models is suggested which include linear, spline and tree models. In addition, combinations such as linear and spline trees are also proposed. Unfortunately, there is no model form amongst these which offers the best fit for all datasets. According to Wolpert any two learning algorithms are equivalent when their performance is averaged across all possible problems (130, 131). This basically means that there is no regression algorithm that outperforms all other regression algorithms across all possible regression datasets. The statement is referred to as the 'No Free Lunch' or 'NFL' theorem in supervised learning. The adage implies the impossibility to get something for nothing, i.e. an algorithm leading to superior model accuracy for all possible datasets. A consequence of the NFL theorem is that the accuracy of a regression algorithm solely depends on the given dataset. Hence, an algorithm may outperform another algorithm on a particular type of dataset but may be inferior to this algorithm on another type of dataset.

In order to build a model to fit the typical non-normal characteristics of LGD data better than a linear model, many studies suggest

alternatives such as tobit models (25, 64), logit models (25, 66), logistic models (61, 63, 65, 68), log-log models (61, 63, 66, 67) or beta models (25, 62, 66, 69). Nonetheless, it is not proven that these significantly fit LGD data better. In addition to distribution characteristics, many meta-learning studies (apart from LGD studies) claim that also other commonly used dataset characteristics such as size, dimensionality, composition and sampling landmarks may favor a specific predictive model algorithm (73, 74, 75, 76, 77, 132, 133, 134). However, the lack of sufficient real-life datasets available to these meta-learning studies (i.e. merely twenty (77) to hundred (78)) undermine the support of these claims.

In spite of the arsenal on meta-learning studies, the current literature is not clear whether and which commonly used dataset characteristics drive regression algorithm fitting performance. In this study, it is explored how simple dataset characteristics may drive the fitting performance of regression algorithms. This may be relevant to support the selection of an optimal model form based on the characteristics of the data to be fit without empirically evaluating each candidate model on the dataset. For this purpose, a meta model is built in order to evaluate how both data based and algorithm based characteristics may favor model accuracy. Data based characteristics involve the number of instances, dichotomous variables, continuous variables and distribution properties of the dependent variable while algorithm based characteristics involve the algorithms performance on very small data samples.

In contrast with previous meta-learning studies, a novel approach is applied here so as to circumvent the scarcity of publicly available regression datasets. The experimental data is constructed by implementing the recently introduced concept of datasetoids (91, 92). A datasetoid is defined as a new dataset obtained by switching an independent variable with a dependent variable. This idea allows to circumvent the scarcity of publicly available real-life datasets (93) by generating more than thousand regression datasetoids to build up a meta dataset. The meta dataset consist of dataset characteristics as independent variables and the performance differences of the considered algorithms as dependent variables. Note that this study covers various real-life model relations, not specific LGD relations, so as to make conclusions towards regression problems in general.

The remainder of this chapter is organized as follows. First, a literature review is conducted on previous meta-learning studies which focus on the algorithm selection based on dataset characterization. Special attention is devoted to the formalization of the algorithm selection problem according to Rice. In the light of Rice's meta-learning framework, the most representing contributions on the subject of meta-learning are further reviewed. Second, the proposed methods to discover possible relations between the characteristics of dataset and the relative accuracy of algorithms are discussed in function of Rice's meta-learning framework. Third, the meta-learning results are discussed and followed by a conclusion.

4.2 Literature review

This section reviews the most important previous studies about the use of meta-learning for algorithm selection. Although the majority of these studies are focused on classification problems, meaningful insights can be extracted for regression problems as well. Both problems only differ in the modeling of the target variable which is discrete in the case of classification and continuous in the case of regression. In order to discuss and compare these studies, they are framed into Rice's abstract model which formalizes the algorithm selection problem (90). Even though Rice's framework does not specify which methods to use, it offers a common language for addressing the components to solve the algorithm selection problem.

4.2.1 Review of Rice's meta-learning framework

Rice's framework for the algorithm selection problem consists of four essential components: the problem space P, the feature space F, the algorithm space A and the performance space Y. Note that these are adapted for the purpose of regression algorithm selection while Rice's framework may cover any kind of algorithm selection. The problem space P is the collection of datasets which consist of a series of values of continuous or dichotomous independent variables and a continuous dependent variable. The algorithm space A is the collection of regression algorithms that can be applied to fit a model to a dataset. The performance space P represents the performance values of a model that is fitted to a dataset with a regression algorithm. The feature space F contains a number of measurable

characteristics for each dataset. The choice of features very much depend on the type of algorithm and should ideally capture all relevant properties of the dataset.

The aforementioned spaces in Rice's framework are connected with each other through mappings. The feature mapping $f : P \rightarrow F$ extracts measurable characteristics from a dataset to a number of features. The selection mapping $s : F \rightarrow A$ chooses the regression algorithm with best performance based on the features extracted from the dataset. The performance mapping $p : P \times A \rightarrow Y$ determines the performance of the regression algorithm applied to the dataset. Hence, the algorithm selection problem can be formally stated as follows: given a dataset $x \in P$ with characteristics $f(x) \in F$, find the algorithm $s(f(x)) \in A$ which maximizes $y \in Y$. An actual algorithm selection tool would thus consist of both the feature mapping f and the selection mapping s combined. Hence, the relevance of such an algorithm selection tool increases when these mappings can be performed more effectively compared to a priori benchmarking experiments.

Based on the above described framework, the following issues need to be addressed to solve the algorithm selection problem. First, a set of regression algorithms that the meta learner can choose from needs to be defined. Second, a set of datasets for both building and validating the meta learner needs to be gathered. Third, a number of dataset features needs to be decided upon so that datasets characterized by similar features correspond to the same algorithms

with similar performance. Forth, the output of the selection mapping needs to be determined so as to provide the actual user recommendation. Fifth, a metric to decide upon the performance of an algorithm applied to a specific dataset problem needs to be determined. Note that the selection algorithm takes as input exclusively the features of the dataset but that the calculation of the performance depends on the original dataset.

4.2.2 Review of previous meta-learning approaches

Rendell and Cho (135) provided one of the earliest contributions to meta-learning by launching the idea that datasets can be characterized by features which could serve as an input for automatic selection mappings or the generation of artificial datasets. Aha (136) used this idea to propose a meta-learning approach for the algorithm selection problem. The suggested features were the number of training instances, the number of classes, the value range, the number of prototypes per class, the relevant and irrelevant attributes, the instance distribution space and the prototype distribution space. Brazdil and Henery (77) extended the study of Aha by incorporating additional features which were also used in a number of subsequent studies (73, 74, 75, 76). These were divided into simple measures (i.e. number of samples, number of attributes, number of classes, number of binary attributes, cost matrix indicator), statistical measures (i.e. standard deviation ratio, mean absolute correlation of attributes, first canonical correlation, fraction separability, skewness, kurtosis) and information theory measures (i.e. entropy of class,

mean entropy of attributes, mean mutual information of class and attributes, equivalent number of attributes, noise signal ratio). The above described features very much focus on the characteristics of the independent and dependent variables separately but it was noticed algorithm performance may however be more depending on the relationship between independent and dependent variables. In addition, it was observed that the computational effort to calculate some features is often greater than for running simple algorithms.

In response, a number of studies were conducted to explore other forms of features which were algorithm based rather than data based. Several studies considered to use properties of specific models as features (e.g. number of nodes and leafs, width and depth of a decision tree) (137, 138, 139) to characterize datasets. Other studies explored the use of relative landmarks (140, 141) and sampling landmarks (132, 133, 134) as features. A relative landmark represents the performance of faster algorithms which may predict the performance of other algorithms. A sampling landmark on the other hand is the performance of an algorithm on a sample of the dataset which may predict the performance of the respective algorithm on the complete dataset. Further, a number of likewise studies were conducted for the purpose of parameter selection rather than algorithm selection. Kuba et al. (142) developed new features for regression specific problems with the aim of selecting parameter settings for SVMs. These included the coefficient of variation, scarcity and stationarity of the dependent variable, presence of outliers, the coefficient of determination of a linear regression model, average abso-

lute correlations between the independent variables themselves and between the independent variables and dependent variable. Soares et al. (143) used these features to build a meta model to select the most optimal width of the Gaussian kernel parameter for SVM regression.

Various types of meta models are proposed to offer the user a recommendation on algorithm selection. This recommendation may be either in the form of a single algorithm or a ranking of the algorithm space. Various studies have used a single algorithm based approach where classification rules are obtained for each algorithm to describe when it significantly outperforms the other algorithms (136) or when it is proven to be applicable for a given dataset (77). Although a single algorithm based approach is most straightforward, the user has no further information about the performance of the other algorithms. In a ranking based approach however, the user might choose a lower ranked algorithm in favor of another criterion as compactness, comprehensibility, computational complexity or familiarity. Some studies presented an instance based learning approach (73, 76) where the most similar dataset in a collection of reference datasets is determined based on some features as described earlier on. The performance of the algorithms on that similar dataset is used to generate a recommendation in the form of a ranking. Other studies provided a recommendation in the form of a ranking by combining pairwise meta models (74, 78). For each pair of algorithms classification rules are induced to indicate whether their accuracy differs significantly or not. Combining these pairwise

meta models altogether in a round robin schedule, a ranking of the algorithm space can be predicted from best to worst accuracy given a particular dataset. In addition, Gamma and Brazdil (144) experimented with regression models to estimate the model error of each algorithm which can be used to form a ranking.

A major problem in meta-learning is the scarcity of publicly available real-life learning problems to build reliable meta models. Two paths may be distinguished to generate extra datasets. A first way is to generate synthetic or artificial datasets (75, 141, 145, 146, 147). The advantage of such an approach is that a finite number of datasets can be generated in order to reliably fit a meta model. The drawback is that it is hard to resemble real-life characteristics. Choices have to be made about the distribution and intercorrelation of the independent variables, and their relation towards the dependent variable (148, 149). Either way, inevitable biases are created this way which are most often undesired (92). A second way to generate datasets is to manipulate existing real-life datasets. This could be done for example by random subsampling (150) with replacement or adding noise to the data. Although these types of generated datasets reflect real-life relations, they also lack sufficient variation or merely hide the same underlying relations a bit more. Another approach however is the use of so called datasetoids, as recently introduced by Soares (91, 92). A datasetoid is defined as a new dataset obtained by switching an independent variable with a dependent variable. Although a datasetoid from a real-life dataset most often does not represent a meaningful learning problem, it

does represent a datasets with relevant real-life relations which is important for the purpose of meta-learning.

As discussed above, a number of different techniques are used to build and validate meta models so as to provide a recommendation to the user on algorithm selection. Although the use of diverse performance metrics, algorithms and features makes it hard to compare predictive performance results, experimental meta models show generally weak performance and are hardly useful in practice. Although the reason for this weak performance could be caused by the use of inappropriate meta algorithms, the lack of predictive power in the meta datasets may rather be the malefactor. Both data based characteristics and algorithm based characteristics are most often used but their predictive power for algorithm selection remains vague. The identification of relevant data based characteristics is a non-trivial task and heavily depends on the algorithm space. The use of sampling landmarks on the contrary is rather straightforward but its additional advantage remains subject for discussion (132, 133, 134). Either way, a major drawback in the past studies is the lack of sufficient real-life datasets to build and test the obtained meta models. The number of meta dataset instances varies from merely 22 to 100 real-life instances which is hardly sufficient for reliable analysis (77, 78). In addition it is noticed that very little meta-learning studies exist on algorithm selection for regression analysis in contrast to classification analysis (151).

4.3 Methods

This section explains the proposed methods so as to implement the components as described in Rice's framework. In concreto, this entails the following issues. First an overview is given which regression techniques are considered to be a candidate to choose from (algorithm space). Next is explained how the accuracy of regression technique are measured in this study (performance mapping). Then is explained which characteristics are extracted from a given dataset as input for the meta models (feature mapping). Further, it is explained how the meta models will decide upon the most accurate regression algorithm based on the dataset characteristics (selection mapping). Finally is clarified how the meta models are developed and tested. Special attention is given to the generation of datasetoids to overcome the scarcity problem of datasets.

4.3.1 Algorithm space

The algorithm space is represented in Table 4.1 and restricted to regression algorithms which produce linear, spline, tree, linear tree and spline tree models. These white box models are considered to be sufficiently comprehensible in order to be applicable for financial institutions real-life LGD modeling. A linear model characterizes the proportional effect of the independent variables individually. A tree model on the other hand is able to represent constant values in different partitions taken into account possible nonlinearities and combined effects of variables. A spline model can be seen as both a generalization of a linear model or a tree model. It extends a linear

model in the sense that it models the data as piecewise linear functions so as to capture nonlinearities. Nonetheless, it also extends a tree model in the sense that it models the different data partitions as linear functions instead of constants. Linear and spline tree models are extensions of tree models which model the leafs as linear or spline functions respectively instead of constants. Although linear and spline tree models aim to fit the data more flexible, their comprehensibility decreases due to increased complexity with respect to ordinary tree models.

Model	**Algorithm**
Linear (L)	OLS is a linear algebra method that builds a multivariate model from linear functions by minimizing the sum of squared residuals.
Spline (S)	MARS is a stepwise method that builds a multivariate model from piecewise linear functions by minimizing the sum of squared residuals.
Tree (T)	CART is a recursive partitioning method that builds a binary decision tree by minimizing the sum of squared residuals.
Linear Tree (LT)	CART/OLS is an extension of the tree regression algorithm which models the leafs as linear models instead of constants.
Spline Tree (ST)	CART/MARS is an extension of the tree regression algorithm which models the leafs as spline models instead of constants.

Table 4.1: Algorithm space

The algorithms employed to build the linear, spline and tree models are Ordinary Least Squares (OLS), Multivariate Adaptive Re-

gression Splines (MARS) and Classification And Regression Trees (CART) respectively (107, 108, 152). The objective of OLS is to find the optimal coefficients of the linear model while CART aims to determine the optimal variables and splits. Seen as a generalization of OLS and CART, the objective of MARS is to select the coefficients as well as the variables and splits. Despite their methodic differences, all three algorithms minimize the sum of squared residuals as criterion. While OLS uses linear algebra, CART and MARS are using an exhaustive search method to solve the regression problem.

OLS can be run without the need for any parameter tuning. To run CART, variance reduction is set to evaluate candidate splitting rules and to determine the optimal depth of the tree. For reasons of comprehensibility, the tree models are restricted to symmetric binary trees with a maximal depth level of four, i.e. maximum sixteen leafs. To run MARS, the penalty for adding a basic function is set to 2.5 as suggested by Hastie et al. (116). Again, for reasons of comprehensibility, the spline models are limited to include first order basic functions only. In contrast to OLS, CART and MARS do implicit variable selection in a recursive and stepwise way respectively. For OLS, an explicit filter method (119) is applied to include important independent variables in the linear models and exclude irrelevant ones by minizing the mean squared error. Any parameter setting or variable selection is performed using a ten times hold out validation schema on the training set.

4.3.2 Performance mapping

The proposed performance mapping consists of calculating the coefficient of determination, denoted R^2, in a hold out validation set up. The coefficient of determination can be defined as one minus the fraction of the residual sum of squares to the total sum of squares:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$$

with y_i the observations, $\overline{y}$ the mean of the observations, $\hat{y}_i$ the predictions and n the number of instances. Since the second term in the formula can be seen as the fraction of unexplained variance, the coefficient of determination can be interpreted as the fraction of explained variance. Although the coefficient of determination can yield negative values when the model predictions are worse than using the mean observations of the training set as prediction (153), it is capped off to zero in these cases in order to obtain strict values between zero and one. A value of zero than refers to a bad data fit less than or equal to that of the mean observations of the training set, while a value of one refers to an excellent data fit where predictions resemble the observations perfectly.

The coefficient of determination as defined above is however susceptible for the phenomenon of statistical shrinkage (154). This implies that adding more independent variables automatically improves the R^2 which may be due to chance alone. In an attempt to take into account this inflation, the adjusted coefficient of determination is

suggested, denoted $\overline{R}^2$ (155):

$$\overline{R}^2 = 1 - (1 - R^2)\frac{n - m}{n - m - 1}$$

with n the number of instances and m the number of independent variables. The adjusted coefficient of determination is a more representative performance metric for the purpose of meta-learning as this involves model comparisons with possibly varying independent variables. Further, the models are validated on a randomly chosen subset that is hold out from the initial dataset. The remaining subset is employed as training set. A rule of thumb is to use about one-fourth for validation and three-fourth for training. To reduce the variability, multiple rounds of hold out validation are performed and averaged. A rule of thumb is to use about ten rounds.

4.3.3 Feature mapping

The proposed feature mapping to characterize datasets is represented in Table 4.2. These include the most popular features found in the meta-learning literature and are adapted to be applied for regression tasks specifically. These are classified into data based features on the one hand and algorithm based features on the other hand. Data based features represent characteristics of the independent variables and dependent variables individually while algorithm based features represent dependency characteristics between both. The data based features cover simple statistics as the number of instances, number of variables and the amount of continuous and dichotomous variables. The size of the dataset is represented by

n being the number of instances in the dataset. In order to characterize the dimensionality of the dataset, the ratio of the number of variables m to the number of instances n is calculated. Further, the continuous composition and dichotomous composition of the dataset is given by the ratio of the number of continuous variables m_c and the number of dichotomous variables m_d to the total number of variables m respectively. In order to characterize the dependent variable, two features are proposed to describe its asymmetry and dispersion. The centrality of the independent variable is defined as the difference between its median y_{50} and mean $\overline{y}$ with respect to its full range $y_{100} - y_0$. A low difference between mean and 50th percentile indicate that the distribution is symmetrical while a large difference refers to asymmetrical distribution. Since it is of no matter whether the distribution is left or right skewed, the absolute value is proposed. The dispersion of the independent variable is represented by its interquartile range $y_{75} - y_{25}$ with respect to its full range $y_{100} - y_0$. A low difference between the 75th and 50th percentile indicate that the distribution is peaked while a large difference refers to a widespread distribution.

The algorithm based features cover dependency characteristics which are represented through sampling landmarks. In this study, sampling landmarks are constructed by calculating the performance of the models under consideration, built and holdout validated on a sample of the dataset. These are assumed to be an indicator of the ten times hold out validation model performance on the complete dataset. More in detail, a sampling landmark is constructed

Feature	Description
Length	The number of instances.
Dimensionality	The ratio of the number of independent variables to the number of instances.
Composition	The ratio of the number of continuous and dichotomous independent variables to the number of independent variables.
Asymmetry	The ratio of the absolute difference between the median and mean of the dependent variable to the full range of the dependent variable.
Dispersion	The ratio of the interquartile range of the dependent variable tot the full range of the dependent variable.
Landmarks	The coefficient of determination of a model built and holdout validated on a random sample of maximum 300 instances.

Table 4.2: Feature mapping

as follows. In first instance, a small but sufficiently representative sample is randomly chosen from the complete dataset. According to Knofczynski and Mundfrom (156), the size of a representative sample for regression analysis depends on the number of independent variables, the chosen model and its corresponding $\overline{R}^2$. Based on this study a sample of about 300 instances on average is sufficiently representative, given an average number of about 13 independent variables and an average $\overline{R}^2$ across all above described models of 47% which is observed in a set of publicly available real-life regression datasets (93). Once a random sample is chosen, the algorithm is holdout validated. The respective model is built by running the corresponding algorithm on a random two-third of the sample. The landmark represents the $\overline{R}^2$, evaluated on the remaining one-third

of the sample. The sampling landmarks are assumed to be directly proportional with the model performance.

4.3.4 Problem space

In order to circumvent the problem of scarcity of real-life datasets, regression datasets with real-life relations may be constructed by employing the recently introduced datasetoid approach (91, 92). For each real-life dataset, a datasetoid is generated as an additional dataset obtained by switching an independent variables with a dependent variable. This way extra learning problems with real-life relations are generated which may not always be meaningful as a learning problem but are all the more relevant for the purpose of meta-learning in particular. Since the size of datasetoids always equals its corresponding original real-life dataset, a subset of each datasetoid is randomly chosen to create variation in the number of instances and variables as this might be a driver for algorithm selection. Further, as datasetoids often do not represent meaningful learning problems, it may occur that there is no relation between its independent variables and dependent variable whatsoever. Because these type of problems are not relevant for the meta-learning study, datasetoids with corresponding $\overline{R^2}$ of zero or less for all considered models are excluded. For the purpose of generating training datasetoids about sixty publicly available real-life classification datasets from different domains are used. Note that classification datasets can just as well be employed when these contain continuous variables, although resulting datasetoids with categorical dependent

variables are excluded. For the purpose of generating test datasetoids about thirty-two publicly available real-life regression datasets are used. Detailed information about both the real-life classification and regression datasets are provided by Alcala et al. (93).

4.3.5 Selection mapping

The proposed selection mapping comprises a set of pairwise regression meta models which aim to predict the performance differences between each pair of models. Notice that it would be convenient to get a recommendation on which type of algorithm is best for this purpose, stressing the relevance of this study. Nonetheless, because the lack of a practical decision support tool, a benchmarking experiment using the above described algorithm space is suggested to select the most appropriate model for meta model construction. Each of the resulting pairwise meta models represent the performance difference of the corresponding two models, denoted $p_{i \to j} = g(f_1, f_2, ..., f_k)$, where $i \neq j = \{L, S, T, LT, ST\}$, $g(.)$ the regression function, $f_1, f_2, ..., f_k$ the features and k the number of features. In order to provide each algorithm with a score, these pairwise performance differences can be combined in a round robin schedule as illustrated in Table 4.3. Doing so, the total performance t_i for each model i is determined by adding its pairwise performance differences. Based on these sums, an additional ranking of the algorithms r_i may be generated. The advantage of predicting pairwise performance differences instead of only predicting which algorithm is best out of a pair, is that not only a ranking can be provided but

that also the size of the performance gaps between algorithms can be quantified. When the performance gap between two algorithms is small, the user might choose a lower ranked algorithm in favor of another criterion such as compactness, comprehensibility, computational complexity or familiarity.

A	L	S	T	LT	ST	t	r
L		$p_{L\to S}$	$p_{L\to T}$	$p_{L\to LT}$	$p_{L\to ST}$	t_L	r_L
S	$p_{S\to L}$		$p_{S\to T}$	$p_{S\to LT}$	$p_{S\to ST}$	t_S	r_S
T	$p_{T\to L}$	$p_{T\to S}$		$p_{T\to LT}$	$p_{T\to ST}$	t_T	r_T
LT	$p_{LT\to L}$	$p_{LT\to S}$	$p_{LT\to T}$		$p_{LT\to ST}$	t_{LT}	r_{LT}
ST	$p_{ST\to L}$	$p_{ST\to S}$	$p_{ST\to T}$	$p_{ST\to LT}$		t_{ST}	r_{ST}

Table 4.3: Selection mapping

4.3.6 Meta model evaluation

The performances of the average based, data based and algorithm based meta model are pairwise compared in order to discover any differences in predictive power between these three set ups. For all ten pairwise meta models in each set up, the predictive performance is determined in terms of $\overline{R}^2$. In order to uncover any significant differences between these performances, a statistical comparison is performed across these three set ups. This is done through a Friedman's test (81) followed by a Holm post-hoc pairwise testing procedure (94) as suggested in the literature (79, 80) for these purposes. Friedman's test is performed to test the null hypothesis that all three set ups perform alike based on the performance of their pairwise meta models. Subsequently, Holm's method is used to compare

each pair of set ups individually.

4.3.7 Implementation details

Standard methods in Matlab are used to build the linear models and regression trees while an external Matlab toolbox is used for building the spline models (ARESLab). Variable selection is performed through the sequential feature selection method in Matlab. For the statistical comparison using Friedman's test and the post-hoc multiple testing procedure of Holm, a stand-alone Java application is used which is provided by Garcia and Herrera (80). All other code required for the experiments is developed by the author.

4.4 Results and discussion

The distributions of both the data based and algorithm based datasetoid features are represented in Figures 4.1 and 4.2 respectively, and the corresponding statistics are shown in Table 4.4. The training meta dataset consists of 680 datasetoid instances (i.e. generated from sixty real-life datasets) and the test meta dataset covers 342 datasetoid instances (i.e. generated from thirty-two real-life datasets). Columns two to twelve represent the aforementioned feature space (i.e. length, dimensionality, continuous composition, dichotomous composition, asymmetry, dispersion, linear, spline, tree, linear tree and spline tree sampling landmark). For all features the mean, the standard deviation, minimum and maximum values are

presented. The training meta data shows similar feature statistics with the test meta data. The number of instances n varies from 350 to 5782 for the training meta data and from 350 to 5815 for the test meta data. Note that these numbers are randomly generated between the interquartile range of what is observed on real-life classification datasets. The smallest datasets are after all considered to be irrelevant and sufficient computer power is lacking for processing the largest datasets. The number of variables m ranges from one to 84 for the training meta data and from one to 85 for the test meta data. The training datasetoids contain 0 to 67 continuous variables and zero to twenty-four dichotomous variables and the test datasetoids are composed of 0 to 68 continuous variables and of 0 to 25 dichotomous variables. The difference between the mean and median of the dependent variables with respect to the total range varies from 0% to 50% for the training meta data and from 0% to 38% for the test meta data. The interquartile range of the dependent variables with respect to the total range goes from 2% to 50% for the training meta data and from 1% to 48% for the test meta data. Sampling landmarks are spread between 0% to 100% in terms of $\overline{R}^2$ for both the training and test meta data.

The distribution of the datasetoid performances are represented in Figure 4.3 and corresponding basic statistics are shown in Table 4.5. The linear, spline, tree, linear tree and spline tree model performances in terms of $\overline{R}^2$ are displayed in the form of histograms for both training and test datasetoids. Based on these histograms, it is clear that the datasetoid model performances a spread out from

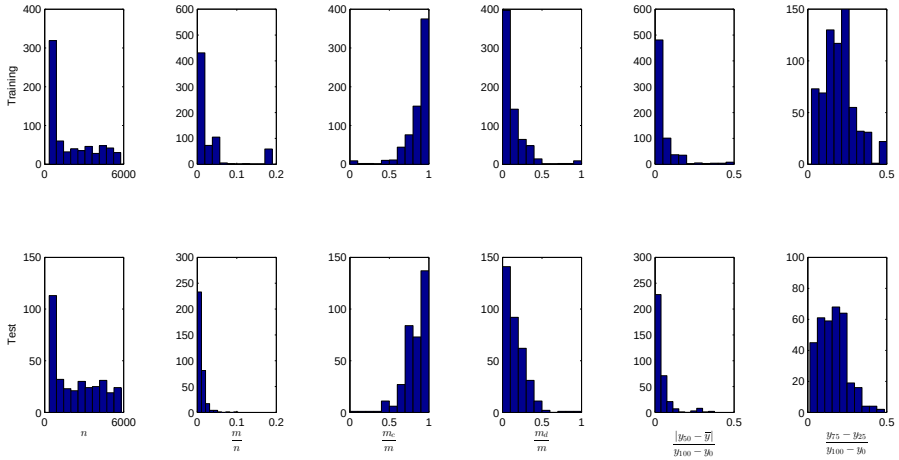

Figure 4.1: Datasetoid data based feature distributions

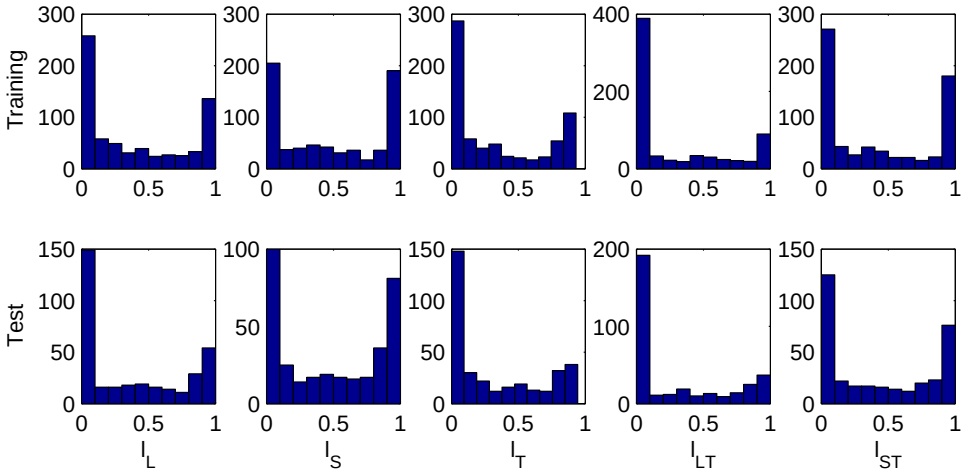

Figure 4.2: Datasetoid algorithm based feature distributions

Statistic	n	$\frac{m}{n}$	$\frac{m_c}{m}$	$\frac{m_d}{m}$	$\frac{\lvert y_{50} - \overline{y} \rvert}{y_{100} - y_0}$	$\frac{y_{75} - y_{25}}{y_{100} - y_0}$	l_L	l_S	l_T	l_{LT}	l_{ST}
Training											
AVG	1913	$3.22{\cdot}10^{-2}$	0.88	0.12	0.05	0.20	0.38	0.48	0.33	0.27	0.41
STD	1738	$5.00{\cdot}10^{-2}$	0.17	0.17	0.08	0.10	0.39	0.40	0.35	0.37	0.41
MIN	350	$1.92{\cdot}10^{-4}$	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00
MAX	5782	$1.90{\cdot}10^{-1}$	1.00	1.00	0.50	0.50	1.00	1.00	0.93	1.00	1.00
Test											
AVG	2369	$9.16{\cdot}10^{-3}$	0.84	0.16	0.04	0.16	0.37	0.48	0.32	0.28	0.43
STD	1739	$1.05{\cdot}10^{-2}$	0.15	0.15	0.06	0.09	0.38	0.39	0.34	0.37	0.40
MIN	350	$4.00{\cdot}10^{-4}$	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00
MAX	5815	$1.00{\cdot}10^{-1}$	1.00	1.00	0.38	0.48	1.00	1.00	0.94	1.00	1.00

Table 4.4: Datasetoid feature statistics

zero to one. This can be seen as a justification that both training and test data possess sufficient variation to cover the total performance range. The histograms also show that the training and test datasetoid performance distributions are very much alike. This can be seen as a justification that the training data is sufficiently representative for the test data. Further, a peak at the first segment and the last segment is dominantly present for all histograms. The peak at the first segment is due to the fact that a majority of the datasetoids lack predictive performance whatsoever. This is not surprising since datasetoids are generally meaningless in real-life. Note that all datasetoids which were not selected a priori even corresponded to a performance equal to or below zero. The peak at the last segment refers to a majority of the datasets with nearly perfect predictive performance. This is most likely caused by heavily correlated independent variables where one of these is shifted into the role of dependent variable during datasetoid fabrication. Although this is seen as a drawback of the use of datasetoids, a representative majority of datasetoids with various performances remains present for reliable meta-learning purposes.

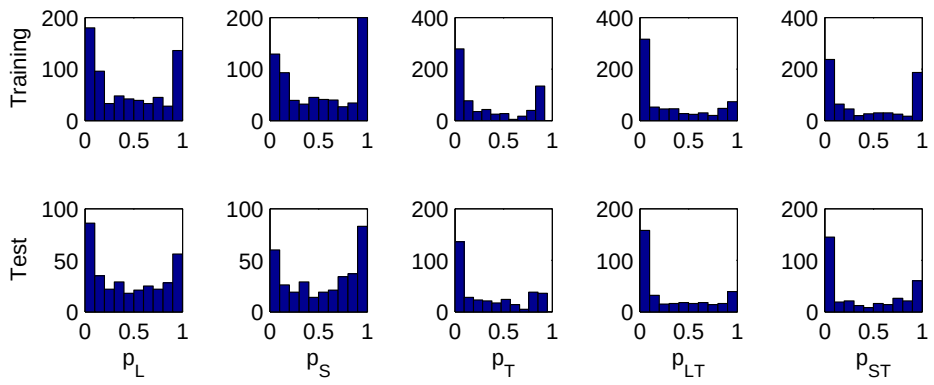

Figure 4.3: Datasetoid model performance distributions

Statistic	p_L	p_S	p_T	p_{LT}	p_{ST}
Training					
AVG	0.44	0.52	0.33	0.31	0.44
STD	0.37	0.38	0.35	0.35	0.41
MIN	0.00	0.00	0.00	0.00	0.00
MAX	1.00	1.00	0.92	1.00	1.00
Test					
AVG	0.45	0.55	0.33	0.31	0.39
STD	0.35	0.36	0.33	0.35	0.39
MIN	0.00	0.00	0.00	0.00	0.00
MAX	1.00	1.00	0.95	1.00	1.00

Table 4.5: Datasetoid model performance statistics

The resulting average based meta model is illustrated in Table 4.6. These can be seen as the benchmark as they represent the most basic meta models. The performance difference between two models is simply predicted by the mean performance difference observed on the training datasetoids. These kind of models are not driven by any dataset features and, hence, correspond to a performance of zero in terms of $\overline{R}^2$. When combining the average based pairwise meta models, the ranked recommendation yields the spline model as best performing model followed by the linear, spline tree, linear

tree and tree model. It can also be seen that the performance of the tree based models are close to each other while linear models and a fortiori spline models show dominantly more performance. When no time is available to predict the best performing model, the average based meta model may offer a general recommendation. These conclusions are at least true on average and are based on the experience on the training meta dataset used in this study. Note that according to the NFL theorem, the performance of these five models should actually be equal on average. The difference can be explained because the meta data used in this study merely covers a small subset (i.e. 680 datasets) out of the infinite amount of all possible datasets for which the NFL theorem holds.

A	L	S	T	LT	ST	t	r
L		-0.08	0.11	0.13	0.00	0.16	2
S	0.08		0.19	0.21	0.08	0.66	1
T	-0.11	-0.19		0.02	0.00	-0.28	5
LT	-0.13	-0.21	-0.02		-0.13	-0.23	4
ST	0.00	-0.08	0.00	-0.13		-0.21	3

Table 4.6: Average based pairwise meta models

The resulting pairwise meta models based on either data features or algorithm features only are displayed in Table 4.7 and Table 4.8 respectively. In order to decide upon which type of meta model form to use, the performances of a linear, spline, tree, linear tree and spline tree are compared across all ten pairwise meta-models. The spline form seems to rank first with an average $\overline{R^2}$ of 27% closely followed by the linear form with an average $\overline{R^2}$ of 25%. Nonetheless,

a linear form is more easy to interpret since a spline form typically includes more terms. Because of its higher comprehensiblity and comparable accuracy with a spline form, a linear form is chosen as meta model form. In order to compare the influence of the features for algorithm selection, their values are adjusted to a notionally common scale. All meta independent variables are standardized by subtracting the meta training dataset mean and dividing this difference with the training meta dataset standard deviation. The mean and standard deviation values of each meta independent variable are reported in Table 4.4.

The data based pairwise meta models in in Table 4.7 seem to imply that these commonly used data characteristics are negligible drivers for algorithm selection. When observing the resulting data based pairwise meta models, it appears that the size and dimensionality of datasets play a role in the prediction of model performance differences because these are included in some pairwise meta models. However, when observing the linear coefficients their influence is rather small. In addition and even more important, the corresponding pairwise meta model performances in terms of $\overline{R^2}$ are barely distinguishable from zero. This means that the meta models with the commonly employed data based features hardly provide any additional advantage compared to the average based meta models. This makes the use of data based features rather irrelevant for this matter.

The algorithm based pairwise meta models in Table 4.8 on the other

Meta model	$\overline{R}^2$
$p_{L\to S} = 0.08 + 0.03 \cdot \frac{m}{n}$	0.00
$p_{L\to T} = -0.11 - 0.05 \cdot n$	0.03
$p_{L\to LT} = -0.13 - 0.04 \cdot \frac{m}{n}$	0.00
$p_{L\to ST} = -0.03 \cdot n + 0.03 \cdot \frac{m}{n} + 0.03 \cdot \frac{m_c}{m}$	0.02
$p_{S\to T} = -0.19 - 0.04 \cdot n$	0.00
$p_{S\to LT} = -0.21 - 0.07 \cdot \frac{m}{n}$	0.00
$p_{S\to ST} = -0.08 - 0.03 \cdot \frac{\lvert y_{50} - \overline{y}\rvert}{y_{100} - y_0}$	0.00
$p_{T\to LT} = 0.06 \cdot n - 0.06 \cdot \frac{m}{n}$	0.04
$p_{T\to ST} = 0.03 \cdot \frac{m}{n} + 0.03 \cdot \frac{m_c}{m}$	0.02
$p_{LT\to ST} = 0.13 + 0.06\frac{m}{n} - 0.03 \cdot \frac{m_d}{m}$	0.04

Table 4.7: Data based pairwise meta models

hand seem to imply that sampling landmarks are important drivers for predicting model performance differences. The coefficients of the sampling landmarks are consistently either negative or positive at the expense or in favor of the corresponding model. The approximately equal weights of the sampling landmarks indicate that the difference between sampling landmarks is directly proportional with model performance. In other words: a single validation run on a random sample of 300 instances seems to be giving an good indication of a ten times hold out validation run on the complete dataset. Note that about half of the datasetoids has a size which is between 350 and 1000 instances. This is close to the sampling landmark size and may cause too optimistic performance results.

Further, it is noticed that the performance of the pairwise meta models are higher when exclusively one stage models (i.e. linear, spline and tree) are involved. These models are after all well dis-

Meta model	$\overline{R}^2$
$p_{L\to S} = 0.08 - 0.10 \cdot l_L + 0.11 \cdot l_S$	0.30
$p_{L\to T} = -0.11 - 0.23 \cdot l_L + 0.24 \cdot l_T$	0.47
$p_{L\to LT} = -0.13 - 0.19 \cdot l_L + 0.17 \cdot l_{LT}$	0.21
$p_{L\to ST} = -0.21 \cdot l_L + 0.26 \cdot l_{ST}$	0.07
$p_{S\to T} = -0.19 - 0.33 \cdot l_S + 0.31 \cdot l_T$	0.56
$p_{S\to LT} = -0.21 - 0.25 \cdot l_S + 0.21 \cdot l_{LT}$	0.41
$p_{S\to ST} = -0.08 - 0.30 \cdot l_S + 0.33 \cdot l_{ST}$	0.12
$p_{T\to LT} = -0.20 \cdot l_T + 0.18 \cdot l_{LT}$	0.40
$p_{T\to ST} = -0.21 \cdot l_T + 0.26 \cdot l_{ST}$	0.07
$p_{LT\to ST} = 0.13 - 0.16 \cdot l_{LT} + 0.23 \cdot l_{ST}$	0.10

Table 4.8: Algorithm based pairwise meta models

tinct from each other, and hence, may be easier to distinguish. Two stage models (i.e. linear tree and spline tree) on the other hand are a combination of the one stage models and are, as such, less distinct from each other, which may explain the lesser performances. Nevertheless, the sampling landmarks as set up in this study do not provide sufficient power to flawlessly predict performance differences. This can however be resolved by adapting the sample size and/or the number of runs to calculate the sampling landmark in function of the dataset size. The latter may be matter for further research.

The performances in terms of $\overline{R}^2$ of the average based, data based and algorithm based meta model are compared in Table 4.9. Table 4.10 illustrates a statistical comparison between these three set ups in order to uncover any significant differences. This is done using the Friedman's test (81) followed by the Holm post-hoc pairwise

testing procedure (94) as suggested in the literature (79, 80). Friedman's test is performed to test the null hypothesis that all three set ups perform alike based on the performance of their pairwise meta models. Subsequently, Holm's method is used to compare each pair of set ups individually. Based on the results, the hypothesis that there is no performance difference between the data based and average based meta model can not be rejected. This means that the use of data based features do not provide any significant additional value above the simple use of the average for algorithm selection which is in contract to most studies in the literature. The hypothesis that there is no performance difference between the algorithm based on the one hand and the data based or average based meta model on the other hand can be rejected with a significance level of 1% and 0% respectively. This means that sampling landmarks are proven to be significantly better drivers than both training averages or data based features for algorithm selection which is rather unclear based on past literature.

4.5 Conclusions

This chapter explores in what degree model performances can be predicted based on both data based and algorithm based characteristics of a given dataset. The study involves experiments with more than thousand datasetoids representing real-life relations, thereby circumventing the scarcity problem of publicly available real-life regression datasets. This study applies the concept of datasetoids to algorithm selection problem which increases the reliability of the

Model	Average based	Data based	Algorithm based
$p_{L\to S}$	0.00	0.00	0.30
$p_{L\to T}$	0.00	0.03	0.47
$p_{L\to LT}$	0.00	0.00	0.21
$p_{L\to ST}$	0.00	0.02	0.07
$p_{S\to T}$	0.00	0.00	0.56
$p_{S\to LT}$	0.00	0.00	0.41
$p_{S\to ST}$	0.00	0.00	0.12
$p_{T\to LT}$	0.00	0.04	0.40
$p_{T\to ST}$	0.00	0.02	0.07
$p_{LT\to ST}$	0.00	0.04	0.10
Average	0.00	0.01	0.27

Table 4.9: Performances of the meta models in terms of $\overline{R}^2$

Hypothesis	z	p
Average based vs algorithm based	3.91	0.00
Data based vs algorithm based	2.80	0.01
Average based vs data based	1.12	0.26

Table 4.10: Statistical comparison of pairwise meta models

results. It is found that data based features such as the size, dimensionality, composition or target distribution of the dataset does not provide any significant additional value above the simple use of the average for algorithm selection which is in contract to several studies in the literature. In addition is proven that sampling landmarks are significantly better drivers than both training averages or data based features for algorithm selection which is rather unclear based on past literature. Although the results of this study are generalizable to other domains as well, these apply in particular for LGD modeling. First, this study presents a selection of algorithms

to build humanly interpretable models. This is important because typical black box models such as those built by Support Vector Machines or Artificial Neural Networks may not be approved by the national regulator because of their lack of comprehensibility, although they may show superior accuracy on various real-life LGD datasets. Second, this study found no evidence that the typical non-normal distribution characteristics of real-life LGD would fit some models better in contrast to some studies who claim the opposite. Consequently, this study advises either to actually compare the model performances for algorithm selection if computationally possible or to compare sampling landmarks as a more time effective way to successfully estimate and compare model performances.

5

Conclusions

"Occurrences in this domain are beyond the reach of exact prediction because of the variety of factors in operation, not because of any lack of order in nature."
-Albert Einstein (German physicist, 1879-1955)

"It is far better to foresee even without certainty than not to foresee at all."
-Henri Pointcare (French mathematician, 1854-1912)

This thesis results in various scientific contributions which are in particular of practical use for financial institutions. First, a benchmarking study is conducted to uncover the predictability of real-life LGD with various types of regression modeling techniques. Second, a backtesting tool is presented to support financial institutions to quantitatively test their internal LGD models. Third, the foundations of a selection tool are established to support model builders to

decide upon the most appropriate model technique. This concluding chapter is organized as follows. First, the most important results of this thesis are briefly summarized. Second, it is explained how these contribute to the scientific literature and how these are important for the industry. Third, several limitations of the conducted studies are highlighted. Fourth and finally, various paths are suggested for further research.

5.1 Results

The first part of this thesis entails a benchmarking study with twenty-four regression techniques and six real-life datasets obtained from major international banking institutions. The average performance of the techniques applied to the real-life LGD datasets ranges from 4% to 43% in terms of R^2. This means that these resulting models have limited explanatory power and thus implies that real-life LGD is hard to predict. Nonetheless, a clear trend can be seen that non-linear techniques yield significantly higher model performance than more traditional linear techniques. This suggests the presence of non-linear relations between the independent variables and the LGD, contrary to a previous benchmarking study on PD modeling where the difference between linear and non-linear techniques is not that explicit. The study clearly demonstrates the potential of non-linear techniques to LGD modeling, possibly in a two stage setting with a linear component so as to improve the comprehensibility of the resulting models.

The second part of this thesis addresses the call for research on backtesting LGD models. Backtesting is a regulatory requirement for any bank implementing the Basel advanced internal ratings based approach. Current backtesting practices most often consist of solely measuring the similarity between model predictions and realized observations. Without proper reference values however, it is not straightforward to determine upon acceptable model performance solely based on these metrics. This study proposes an workbench of statistical hypothesis tests which includes standard parametric and non-parametric tests as well as a number of non-standard tests constructed through a bootstrapping approach based on commonly used LGD model performance metrics. These tests are applied in such a way that they take into account an appropriate reference value indicating acceptable accuracy in addition to the number of LGD observations.

The third part of this thesis explores in what degree model performance differences can be predicted based on both data based and algorithm based characteristics of a given dataset. The study involves experiments with more than thousand datasetoids representing real-life relations, thereby circumventing the scarcity problem of publicly available real-life regression datasets. It is found that data based features such as the size, dimensionality, composition or target distribution of the dataset do not provide any significant additional value above the simple use of the average for algorithm selection which is in contrast to most studies in the literature. In addition is proven that sampling landmarks are significantly better

drivers than both training averages or data based features for algorithm selection which is rather unclear based on past literature.

5.2 Contributions

The benchmarking study is the first large scale LGD study in terms of both regression techniques and real-life LGD datasets. Its value is in particular in the use of default data from major international banks. It is not straightforward to obtain real-life LGD data for research purposes because financial institutions either have no sufficiently large track record of losses at the moment or choose not to share these for scientific research because of reasons of confidentiality. The results of this study may offer other financial institutions valuable information about the performance of techniques on real-life LGD. In first instance, this study can help banks in selecting the appropriate regression algorithm to model their LGD portfolio's. In second instance, banks are provided with an indication of the performance of LGD models. This knowledge can serve to validate their own internal LGD models by comparing these with the model performances acquired in this study.

The backtesting study is the first study to introduce how statistical hypothesis tests can be applied for validating LGD models. The importance of validation methods for LGD models is increased since the deployment of the advanced internal ratings based approach. Although Basel requires banks to validate their internal LGD models at least yearly, it does not further specify how to perform this

validation. This study may fill in this gap. The proposed statistical hypothesis tests may be valuable for financial institutions implementing the advanced internal ratings based approach. In order to be Basel compliant, banks are required to have a documented approach towards the validation of their internal LGD models. Banks can implement the proposed LGD backtesting methods and refer to this study to prove the soundness of both their internal LGD models and their validation process to the national regulator.

The meta-learning study is the first study to apply the concept of datasetoids to algorithm selection which increases the reliability of the results. The study proves that data based features do not matter while sampling landmarks do matter for algorithm selection. These findings are either in contrast with the literature or rather unclear based on previous studies. The results of this study may be of practical use when models need to be fit on large datasets. In these cases, sampling landmarks can be a time saving way to either select the most accurate model technique or to optimize a specific model parameter. Financial institutions in particular may notice the average superiority of spline models for comprehensible regression analysis. In addition to the results of the LGD benchmarking study, the fitting of splines to LGD data may be preferred to develop both an accurate and comprehensible LGD model.

5.3 Limitations

Although the benchmarking study is up to now the largest LGD study in terms of the number of datasets and techniques, it encounters limitations from a statistical point of view in two ways. First, the number of real-life LGD datasets available for this research is rather low. The low number of datasets causes to decrease the statistical power of the hypothesis tests to detect significant differences between the considered algorithms when there actually are. Second, all algorithms are merely single hold out validated because of reasons of computational complexity. A single hold out schema may distort the resulting model performances because of the large variation on these results when compared to a multiple hold out schema (e.g. ten fold cross validation).

Although the backtesting study results in a workbench of statistical hypothesis test to quantitatively evaluate the performance of LGD models, these tests may be rather powerless for low default portfolios. Several types of portfolios can be characterized by a low number of defaults. These typically include portfolios of exposures to sovereigns, large banks or insurance companies. Other examples are recent market entrants for a given portfolio or portfolios with long workout periods. When the LGD portfolio lacks sufficient observations, the statistical power of the corresponding statistical hypothesis tests decrease. This means the concerning tests are not able to detect significant model deterioration, even if this actually would be the case. Hence, the use of the proposed statistical hypothesis tests is limited to portfolios with sufficient defaults.

Although the meta-learning study slightly circumvented the lack of sufficient real-life datasets by extracting a multiple of so called datasetoids, this method involves the following drawbacks. About half of the generated datasetoids are less representable because there predictive power either is extreme low or extreme high. On the one hand datasetoids are generally meaningless content wise and often may result datasetoids lacking any predictive performance at all. On the other hand, datasetoids based on heavily correlated independent variables where one of these is shifted into the role of dependent variable during datasetoid fabrication result in nearly perfect predictive performance. An additional limitation is the restricted sample size of both datasetoids and sampling landmarks. Because of reasons of computational complexity, large dataset sizes (i.e. more than about 6000) are not represented in this study. In addition about half of the datasetoids has a size close to the sampling landmark size (i.e. between 350 and 1000) and may cause too optimistic performance results. Further, the size of the sampling landmarks is constant while ideally should be depending on size of the datasetoid size.

5.4 Future research

Although the benchmarking study proves that there are significant performance differences between algorithms, the observed predictive model performances are considered low in general. This may be an indication that real-life LGD datasets are lacking predictive power. A possible path for future research could exist to search

for variables with higher predictive power. Such a study can be conducted via a Delphi method which relies on a panel of interacting and anonymous LGD model experts. This panel of experts can be given the task to identify and to prioritize the high predictive candidate LGD drivers for various types of portfolios. In addition the panel can be asked to agree upon a predictive model involving the earlier identified drivers in an attempt to construct an expert based LGD model. Subsequently, this expert based model can be validated in two ways. First, the expert based model can be backtested by comparing its predictions with actual loss observations. Second, the predictive performance of the expert based model can be benchmarked by comparing them with empirically built models.

Although the backtesting study provides various useful tests to quantitatively validate the predictive performance of LGD models, these are rather useless for low default portfolios. Hence, there is a need to develop applicable procedures to validate LGD models for portfolios which are characterized by a small amount of observations. Since banks and supervisors may find that backtesting LGDs for low default portfolios can not be done in a way that strongly demonstrates good predictiveness in a quantitative way, more emphasis may be on qualitative techniques in order to satisfy both themselves and their supervisor that their predictions are reasonable. A study to extract best practices for qualitative validation techniques could offer relief. The focus may be rather on the process of collecting data, designing the predictive model and how it is used in daily practice.

The meta-learning study can be extended in a number of ways in order to create a practical support tool for algorithm selection. First, it seems that a single validation run on a sample of 300 instances is not always sufficiently representative for a ten times hold out validation on the full dataset. This problem could be addressed with a study aiming to build a model to estimate a sampling landmark's sample size in function of the required statistical power, the full dataset size and the required model. Second, the relevance of an algorithm selection model increases when the time effort to conduct a benchmarking study is too high. The gained time profit when using sampling landmarks for selecting purposes may be studied in function of the size of the full dataset, the type of algorithms and their run times. Third, the comprehensibility of a model is often labeled as very important but rather difficult to measure objectively across different types of models. Therefore, an empirical study may be set up with the goal of measuring the comprehensibility of models on a common scale.

Appendix A

Specification of the risk weight function

The internal ratings based (IRB) approach (5) as introduced in Basel II allows banks to use their own internal measures for key drivers of credit risk as primary inputs to the capital calculation, subject to meeting certain conditions and to explicit supervisory approval. All institutions using the IRB approach are allowed to determine the borrowers probabilities of default while those using the advanced IRB approach will also be permitted to rely on own estimates of loss given default and exposure at default on an exposure-by-exposure basis. These risk measures are converted into risk weights and regulatory capital requirements by means of a risk weight formula specified by the Basel Committee. This section describes the economic foundations, the regulatory requirements as well as the underlying mathematical model of the IRB approach.

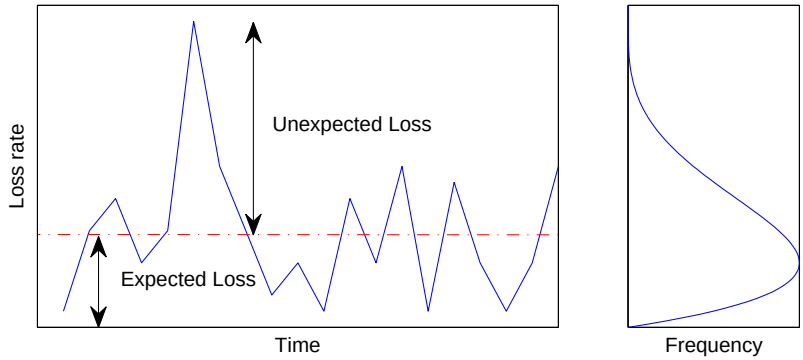

Figure A.1: Loss rate (5)

Economic foundation

The occurrence of loss of interest and principal capital is inherently connected to the credit business. Because of the credit risk arising from borrowers who do not make payments as promised, defaults happen to occur. The number and severity of defaults can vary from year to year. An example of realized losses in a particular portfolio over time is captured in Figure A.1. The variation of these losses can be illustrated with the portfolio loss distribution as seen in Figure A.2. A distinction is made between expected loss (EL) and unexpected loss (UL).

The expected loss of a portfolio is the average level of credit loss a bank can reasonably expect to experience. This is seen as the normal cost of doing business. It is mainly covered by the interest rate charged to the obligors and by provisioning. The expected loss gives information about the location of the portfolio loss distribution.

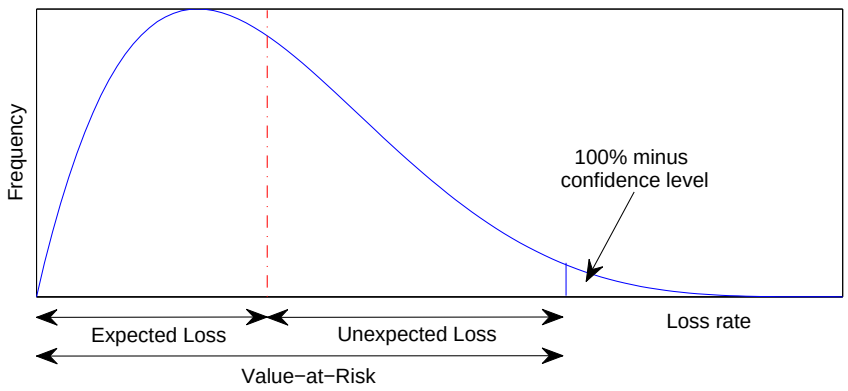

Figure A.2: Loss distribution (5)

The unexpected loss of a portfolio is the loss that differs from the expected loss. Financial institutions know they will happen to occur now and then but do not know in advance their timing and severity. Banks need buffer capital to absorb these peak losses if they occur so as to protect their own obligations. The unexpected loss gives information about the dispersion of the portfolio loss distribution.

The value at risk (VaR) of a portfolio is the sum of the expected loss and the unexpected loss that a bank is able to cover through both profits and capital respectively. The value at risk is defined at a given confidence level. The latter is the likelihood that a bank will remain solvent when capital is set according to the unexpected loss gap and if the expected loss is covered by provisions and revenues. The value at risk is the corresponding threshold for a given confidence level.

Regulatory requirements

The Basel IRB model used for the derivation of supervisory capital charges for unexpected loss is subject to an important restriction in order to fit supervisory needs. The model should be portfolio invariant, i.e. the capital required for any given loan should only depend on the risk of that loan and must not depend on the portfolio it is added to (5). This regulatory requirement is set forth for reasons of simplicity. For supervisory needs, it is generally too complex to take into account the composition of the portfolio to determine the required capital for each single loan. It can be shown that only so-called Asymptotic Single Risk Factor (ASRF) models are portfolio invariant (157).

ASRF models assume that a) the portfolio is asymptotically fine-grained and b) that there is only one single systematic risk factor. When a portfolio consists of a large number of relatively small exposures, idiosyncratic risks associated with individual exposures tend to cancel out one-another and only systematic risks that affect many exposures have a material effect on portfolio losses (5). Note that, although the use of ASRF models is suggested, Basel does by no means enforce banks to employ a specific model. Due to the portfolio invariance property, regulatory capital depend only on the characteristics of the obligor and not on the characteristics of the remainder of the portfolio (157). As such, obligor specific attributes like the probability of default (PD), the loss given default (LGD) and the exposure at default (EAD) suffice to determine the capital charges.

Model specification

So far, the unexpected loss for which banks should hold capital as a safety cushion has been regarded from a top-down perspective, i.e. as the difference of the value-at-risk and the expected loss of the portfolio loss distribution:

$$UL = VaR - EL$$

In what follows, the unexpected loss is built from the bottom-up, namely from its components PD, LGD and EAD. This eventually leads to the Basel IRB formula to determine the regulatory capital to cover the estimated unexpected loss:

$$\begin{aligned} RC = UL &= 12.5 \cdot \sum_{i=1}^{N} \overline{EAD^*}_i \cdot \overline{LGD^*}_i \\ &\cdot \left[\Phi_N \left[\sqrt{\frac{1}{1-\rho}} \phi_N^{-1}(PD_i) + \sqrt{\frac{\rho}{1-\rho}} \Phi_N(0.999) \right] - PD_i \right] \\ &\cdot \left(\frac{1 + (M_i - 2.5) \cdot b(PD_i)}{1 - 1.5 \cdot b(PD_i)} \right) \end{aligned}$$

where the factor 12.5 is introduced so as to fit the 8% capital adequacy rule, i.e. $12.5 \cdot 0.08 = 1$.

Expected loss

The expected loss of a portfolio is the sum of the expected loss of each single loan in the portfolio:

$$EL = E[L_P] = \sum_{i=1}^{N} E[L_i]$$

where L_P and L_i denote the loss of the entire portfolio P and the loss of an individual loan i respectively. The expected loss of an individual loan is a stochastic variable and is assumed to follow the equation below:

$$L_i = EAD_i \cdot LGD_i \cdot \delta_{PD_i}$$

where δ_{PD} is either 0 (non-default) or 1 (default). Hence, the expected value of the portfolio loss equals:

$$E[L_P] = \sum_{i=1}^{N} \overline{EAD}_i \cdot \overline{LGD}_i \cdot PD_i$$

where $\overline{EAD}$ and $\overline{LGD}$ denote the average exposure at default and loss given default respectively.

Value at risk

The value-at-risk is the level of capital that is required to prevent the bank from going bankrupt in one year with a probability of no more than 100% minus the confidence level. For this purpose, Vasicek's model was adopted as the heart of Basel's IRB formula (158):

$$VaR_i(\alpha) = \sum_{i=1}^{N} \overline{EAD}_i \cdot \overline{LGD}_i \cdot \Phi_N \left[\sqrt{\frac{1}{1-\rho}} \Phi_N^{-1}(PD_i) + \sqrt{\frac{\rho}{1-\rho}} \Phi_N^{-1}(\alpha) \right]$$

where Φ the cumulative standard normal distribution, α is the confidence interval and ρ is the asset correlation. The Vasicek formula is derived from an adaptation of Merton's credit risk model (159). However, Merton is interested in the value of equity of a single firm in isolation, whereas Vasicek is interested in the probability of default on portfolio debt of a bank (160).

Vasicek's formula is used to determine an appropriate downturn PD, i.e. the conditional PD given economic downturn conditions. In a first step, the default threshold for the PD is determined by applying the inverse cumulative standard normal distribution function to the PD. Likewise, a default threshold for an appropriate conservative value of the single risk factor can be derived by applying the inverse cumulative standard normal distribution function to the predetermined supervisory confidence level. A correlation-weighted sum of both the default threshold and the conservative value of the single risk factor yields a downturn default threshold. In a second step, the downturn PD is determined by applying the cumulative standard normal distribution function to the downturn default threshold.

Confidence level

The supervisory confidence level α is fixed at 99.9%. This means a bank will not have sufficient capital to cover its losses in 1 out of 1000 years. A capital cushion with $\alpha = 0.999$ would be far in excess of most regulators' actual requirements if the Vasicek formula's

assumptions approximated reality. The high confidence level is justified so as to provide an appropriate conservative value of the single risk factor, given the Vasicek model uncertanties. Estimation errors might inevitable occur from banks' internal PD, LGD and EAD estimation (5). Further, Vasicek assumes an infinitely fine-grained portfolio and a normally distributed single risk factor which is rarely the case in reality (160).

Asset correlation

The asset correlation ρ shows how the asset value of one borrower depends on the asset value of another borrower. In an ASRF model all borrowers are linked to each other by the systematic risk factor that can be interpreted as a reflection of the state of the global economy. Hence, the asset correlation may also be expressed as the degree of the obligor's exposure to that systematic risk factor. The higher the asset correlation the more likely becomes higher unexpected losses. This means portfolios with higher asset correlations require bigger capital cushions. The asset correlation is empirically derived with different approaches and results for corporate exposures on the one hand and retail exposures on the other hand.

The supervisory asset correlations for corporate exposures have been derived by the analysis of datasets from G10 supervisors. The analysis of these time series revealed that asset correlations decrease

with increasing PD and firm size S:

$$\rho = 0.12 \cdot \frac{1 - e^{-50 \cdot PD}}{1 - e^{50}} + 0.24 \cdot \left(1 - \frac{1 - e^{-50 \cdot PD}}{1 - e^{50}}\right) - 0.04 \cdot H \cdot (1 - \frac{S - 5}{45})$$

with values ranging from 12% to 24%, the correlations decreasing with a pace of 50 and where the size adjustment factor affects borrowers with annual sales between 5 million and 50 million ($H = 1$ for $S \leq 5$ and $H = 0$ for $S \geq 50$). Note that the asset correlation for bank and sovereign exposures is the same as for corporate borrowers, only that the size adjustment factor does not apply.

The asset correlations for retail exposures have been reverse engineered from economic capital figures from large internationally active banks and historical loss data from supervisory databases of the G10 countries. This led to three different correlation functions: a relatively high and constant correlation of $\rho = 0.15$ for residential mortgages, a relatively low and constant correlation of $\rho = 0.04$ for qualifying revolving retail exposures and a PD dependent correlation for other retail exposures:

$$\rho = 0.03 \cdot \frac{1 - e^{-35 \cdot PD}}{1 - e^{35}} + 0.16 \cdot \left(1 - \frac{1 - e^{-35 \cdot PD}}{1 - e^{35}}\right)$$

The latter is structurally equivalent to the corporate correlation function. However, its lowest and highest values range from 3% to 16% and the correlations decrease with a pace of 35.

Maturity

Portfolios consist of loans with different maturities. Both intuition and empirical evidence indicate that long-term credits are riskier than short-term credits (5). Hence, the required capital should increase with maturity. Since the Vasicek formula calculates capital for a one year horizon, the IRB formula is adjusted for loans with a maturity over one year. The Basel maturity adjustment M_{adj} is derived by applying a specific market-to-market credit risk model to capture the time structure of PD (i.e. the likelihood and magnitude of PD changes) which leads to:

$$M_{adj} = \frac{1 + (M - 2.5) \cdot b(PD)}{1 - 1.5 \cdot b(PD)}$$

with $b(PD) = (0.11852 - 0.05478 \cdot ln(PD))^2$. The maturity adjustment increases linearly with M and decreases with an increasing PD. Note that the maturity adjustment is only applicable for sovereign, bank and corporate exposures, but not for retail exposures as the asset correlation for retail implicitly contains maturity effects due to its empirical derivation (5).

Downturn LGD and EAD

The LGD parameter used to calculate the unexpected loss must also reflect adverse economic scenario's (5). During an economic downturn typically higher losses are reported than under normal business conditions (48, 161, 162, 163). Therefore, Basel requires banks to use their own estimate of downturn loss given default $\overline{LGD^*}$ instead

of a supervisory function to map $\overline{LGD}$ to $\overline{LGD^*}$. Because LGD estimation is a new and emerging field, the Basel committee determined that it would be inappropriate to apply a single supervisory LGD mapping function (as opposed to PD) (5). Likewise, a downturn exposure at default $\overline{EAD^*}$ is required for the calculation of the unexpected loss. Note that Basel decided to also use downturn loss given default $\overline{LGD^*}$ and downturn exposure at default $\overline{EAD^*}$ for the calculation of EL. However this results in a higher expected loss as LGD^*s and EAD^*s are generally higher, an additional compliance and validation burden is avoided that would arise if banks were required to estimate and report both $\overline{LGD^*}$ and $\overline{LGD}$s for the exposures (5).

Appendix B

Results of the benchmarking experiment

Technique	**MAE**	**RMSE**	**AUC**	**AOC**	R^2	r	ρ	τ
OLS	0.3257	0.3716	0.6570	0.1380	0.0972	0.3112	0.3084	0.2145
B-OLS	0.3474	0.4294	0.6580	0.1843	-0.2060	0.2954	0.2991	0.2071
BR	0.3356	0.3693	0.5690	0.1363	0.0546	0.2601	0.2641	0.1844
BC-OLS	0.3835	0.4579	0.5180	0.2096	-0.3747	0.2403	0.2312	0.1602
RiR	0.3267	0.3723	0.6561	0.1385	0.0933	0.3056	0.3033	0.2106
RoR	0.3262	0.3723	0.6565	0.1385	0.0935	0.3061	0.3034	0.2107
RT	0.3228	0.3732	0.5990	0.1392	0.0892	0.2997	0.2913	0.2095
MARS	0.3214	0.3704	0.6657	0.1372	0.1027	0.3205	0.3122	0.2187
LSSVM	0.3184	0.3669	0.6723	0.1346	0.1194	0.3466	0.3442	0.2444
ANN	0.3118	0.3648	0.6840	0.1331	0.1295	0.3603	0.3559	0.2524
LOG+OLS	0.3202	0.3700	0.6210	0.1366	0.1063	0.3262	0.3143	0.2214
LOG+B-OLS	0.3163	0.3750	0.6020	0.1406	0.1002	0.3166	0.3103	0.2185
LOG+BR	0.3560	0.4142	0.5270	0.1715	0.0782	0.2797	0.2591	0.1794
LOG+BC-OLS	0.4308	0.5090	0.5040	0.2590	-0.6946	0.2125	0.2440	0.1731
LOG+RiR	0.3193	0.3693	0.6655	0.1363	0.1081	0.3289	0.3167	0.2234
LOG+RoR	0.3171	0.3700	0.6554	0.1369	0.1045	0.3264	0.3205	0.2270
LOG+RT	0.3219	0.3693	0.6160	0.1363	0.1081	0.3301	0.3212	0.2263
LOG+MARS	0.3205	0.3689	0.6658	0.1360	0.1099	0.3320	0.3248	0.2286
LOG+LSSVM	0.3191	0.3679	0.6664	0.1353	0.1150	0.3401	0.3336	0.2371
LOG+ANN	0.3174	0.3664	0.6320	0.1342	0.1221	0.3502	0.3406	0.2395
OLS+RT	0.3170	0.3681	0.6730	0.1354	0.1137	0.3382	0.3342	0.2348
OLS+MARS	0.3177	0.3679	0.6799	0.1353	0.1150	0.3394	0.3363	0.2352
OLS+LSSVM	0.3115	0.3631	0.6929	0.1317	0.1379	0.3714	0.3666	0.2596
OLS+ANN	0.3079	0.3633	0.6960	0.1318	0.1367	0.3716	0.3638	0.2581

Table B.1: BANK1 model performance results

Technique	**MAE**	**RMSE**	**AUC**	**AOC**	R^2	r	ρ	τ
OLS	0.1187	0.1613	0.8100	0.0259	0.2353	0.4851	0.4890	0.3823
B-OLS	0.1058	0.1621	0.8000	0.0262	0.2273	0.4768	0.4967	0.3881
BR	0.1020	0.1661	0.7300	0.0275	0.2120	0.4635	0.4857	0.3861
BC-OLS	0.1056	0.1623	0.7450	0.0262	0.2226	0.4718	0.4990	0.3900
RiR	0.1187	0.1606	0.8074	0.0258	0.2415	0.4915	0.4855	0.3792
RoR	0.1075	0.1663	0.8063	0.0277	0.1866	0.4770	0.4824	0.3751
RT	0.0978	0.1499	0.7710	0.0224	0.3390	0.5823	0.5452	0.4357
MARS	0.1068	0.1531	0.8397	0.0234	0.3113	0.5579	0.5321	0.4168
LSSVM	0.1047	0.1518	0.8365	0.0230	0.3229	0.5690	0.5301	0.4160
ANN	0.0956	0.1472	0.8530	0.0216	0.3632	0.6029	0.5549	0.4366
LOG+OLS	0.1060	0.1622	0.7590	0.0255	0.2268	0.4838	0.5206	0.4084
LOG+B-OLS	0.1040	0.1567	0.8320	0.0245	0.2779	0.5286	0.5202	0.4083
LOG+BR	0.1015	0.1688	0.7250	0.0285	0.2024	0.4529	0.4732	0.3876
LOG+BC-OLS	0.1034	0.1655	0.7320	0.0273	0.2124	0.4628	0.4870	0.3820
LOG+RiR	0.1049	0.1554	0.8312	0.0240	0.2901	0.5386	0.5209	0.4091
LOG+RoR	0.1043	0.1558	0.8307	0.0242	0.2859	0.5350	0.5200	0.4084
LOG+RT	0.1041	0.1538	0.8360	0.0236	0.3049	0.5545	0.5254	0.4126
LOG+MARS	0.1031	0.1537	0.8355	0.0236	0.3059	0.5531	0.5268	0.4149
LOG+LSSVM	0.1031	0.1530	0.8334	0.0234	0.3121	0.5587	0.5243	0.4128
LOG+ANN	0.1011	0.1531	0.8430	0.0234	0.3109	0.5585	0.5380	0.4240
OLS+RT	0.1015	0.1506	0.8410	0.0227	0.3331	0.5786	0.5344	0.4188
OLS+MARS	0.1081	0.1526	0.8379	0.0233	0.3150	0.5615	0.5300	0.4156
OLS+LSSVM	0.1029	0.1520	0.8428	0.0230	0.3208	0.5665	0.5398	0.4241
OLS+ANN	0.0999	0.1474	0.8560	0.0217	0.3612	0.6010	0.5585	0.4398

Table B.2: BANK2 model performance results

Technique	**MAE**	**RMSE**	**AUC**	**AOC**	R^2	r	ρ	τ
OLS	0.0549	0.1411	0.6460	0.0178	0.0124	0.1168	0.0965	0.0718
B-OLS	0.0348	0.1449	0.6610	0.0188	-0.0419	0.0767	0.1754	0.1361
BR	0.0883	0.1315	0.6530	0.0169	-0.1128	0.1567	0.1719	0.1323
BC-OLS	0.0340	0.1456	0.6380	0.0190	-0.0529	0.1373	0.2312	0.1765
RiR	0.0550	0.1405	0.6499	0.0177	0.0210	0.1460	0.1270	0.0936
RoR	0.0347	0.1453	0.6438	0.0189	-0.0464	0.1733	0.1991	0.1501
RT	0.0482	0.1311	0.6990	0.0154	0.1477	0.3869	0.2007	0.1673
MARS	0.0478	0.1229	0.7345	0.0131	0.2506	0.5016	0.1344	0.0974
LSSVM	0.0473	0.1270	0.7441	0.0140	0.1998	0.4526	0.2085	0.1520
ANN	0.0458	0.1318	0.6000	0.0152	0.1386	0.3776	0.1482	0.1105
LOG+OLS	0.0553	0.1417	0.6010	0.0179	0.0043	0.0759	0.0701	0.0510
LOG+B-OLS	0.0392	0.1429	0.6330	0.0182	-0.0127	0.1214	0.1252	0.0923
LOG+BR	0.0569	0.1417	0.5790	0.0180	0.0043	0.0742	0.1710	0.1265
LOG+BC-OLS	0.0349	0.1448	0.6330	0.0188	-0.0395	0.1665	0.1918	0.1426
LOG+RiR	0.0545	0.1408	0.6404	0.0177	0.0169	0.1319	0.1511	0.1094
LOG+RoR	0.0366	0.1440	0.6510	0.0185	-0.0277	0.1510	0.2057	0.1504
LOG+RT	0.0434	0.1297	0.7210	0.0146	0.1663	0.4553	0.1571	0.1170
LOG+MARS	0.0467	0.1264	0.7365	0.0139	0.2082	0.4884	0.1381	0.0998
LOG+LSSVM	0.0460	0.1312	0.7485	0.0151	0.1471	0.4152	0.2272	0.1676
LOG+ANN	0.0452	0.1219	0.6190	0.0133	0.2634	0.5381	0.1671	0.1242
OLS+RT	0.0540	0.1372	0.7050	0.0168	0.0660	0.2578	0.1748	0.1285
OLS+MARS	0.0471	0.1229	0.7189	0.0131	0.2512	0.5018	0.1231	0.0879
OLS+LSSVM	0.0483	0.1258	0.7416	0.0137	0.2148	0.4648	0.1869	0.1354
OLS+ANN	0.0570	0.1388	0.6730	0.0171	0.0442	0.2605	0.1369	0.1005

Table B.3: BANK3 model performance results

Technique	**MAE**	**RMSE**	**AUC**	**AOC**	R^2	r	ρ	τ
OLS	0.2712	0.3479	0.8520	0.1208	0.4412	0.6643	0.5835	0.4331
B-OLS	0.2214	0.3743	0.8500	0.1396	0.3530	0.6510	0.5822	0.4321
BR	0.3208	0.3777	0.8480	0.1425	0.3405	0.6527	0.5908	0.4452
BC-OLS	0.3185	0.4292	0.6750	0.1839	0.1478	0.5726	0.5820	0.4316
RiR	0.2707	0.3473	0.8541	0.1204	0.4429	0.6657	0.5972	0.4495
RoR	0.2576	0.3607	0.8483	0.1299	0.3992	0.6527	0.5857	0.4402
RT	0.2476	0.3362	0.8480	0.1128	0.4782	0.6916	0.5919	0.4762
MARS	0.2617	0.3361	0.8636	0.1128	0.4783	0.6917	0.6162	0.4631
LSSVM	0.2428	0.3315	0.8655	0.1097	0.4924	0.7017	0.6203	0.4692
ANN	0.2393	0.3299	0.8670	0.1086	0.4974	0.7053	0.6109	0.4555
LOG+OLS	0.2577	0.3465	0.8520	0.1199	0.4455	0.6678	0.5840	0.4338
LOG+B-OLS	0.2399	0.3551	0.8500	0.1259	0.4176	0.6651	0.5801	0.4301
LOG+BR	0.2738	0.3560	0.8520	0.1265	0.4147	0.6680	0.5868	0.4342
LOG+BC-OLS	0.2502	0.3489	0.8510	0.1215	0.4379	0.6659	0.5819	0.4322
LOG+RiR	0.2538	0.3432	0.8572	0.1176	0.4559	0.6755	0.6026	0.4543
LOG+RoR	0.2354	0.3521	0.8534	0.1238	0.4275	0.6728	0.5960	0.4477
LOG+RT	0.2679	0.3621	0.8570	0.1309	0.3945	0.6656	0.5899	0.4364
LOG+MARS	0.2536	0.3433	0.8572	0.1177	0.4558	0.6754	0.6027	0.4544
LOG+LSSVM	0.2534	0.3425	0.8590	0.1172	0.4581	0.6771	0.6024	0.4541
LOG+ANN	0.2558	0.3457	0.8540	0.1184	0.4480	0.6698	0.5852	0.4348
OLS+RT	0.2628	0.3425	0.8590	0.1171	0.4582	0.6776	0.6017	0.4498
OLS+MARS	0.2617	0.3362	0.8620	0.1128	0.4781	0.6915	0.6117	0.4582
OLS+LSSVM	0.2439	0.3322	0.8656	0.1102	0.4904	0.7003	0.6211	0.4698
OLS+ANN	0.2404	0.3300	0.8710	0.1087	0.4971	0.7053	0.6195	0.4635

Table B.4: BANK4 model performance results

Technique	**MAE**	**RMSE**	**AUC**	**AOC**	R^2	r	ρ	τ
OLS	0.1875	0.2375	0.7480	0.0555	0.2218	0.4740	0.5192	0.3651
B-OLS	0.1861	0.2368	0.7410	0.0561	0.2263	0.5073	0.5168	0.3636
BR	0.1957	0.2402	0.7240	0.0575	0.2038	0.4557	0.4811	0.3359
BC-OLS	0.1848	0.2373	0.7390	0.0560	0.2228	0.5014	0.5155	0.3632
RiR	0.1864	0.2373	0.7467	0.0555	0.2233	0.4775	0.5238	0.3704
RoR	0.1892	0.2430	0.7406	0.0579	0.1852	0.4543	0.5121	0.3612
RT	0.1851	0.2324	0.7370	0.0538	0.2546	0.5056	0.4957	0.3888
MARS	0.1733	0.2222	0.7709	0.0488	0.3187	0.5666	0.5565	0.3980
LSSVM	0.1707	0.2198	0.7847	0.0479	0.3331	0.5794	0.5801	0.4167
ANN	0.1678	0.2173	0.7830	0.0470	0.3486	0.5964	0.5765	0.4148
LOG+OLS	0.1851	0.2336	0.7500	0.0542	0.2468	0.4975	0.5246	0.3704
LOG+B-OLS	0.1852	0.2347	0.7480	0.0548	0.2397	0.5117	0.5192	0.3658
LOG+BR	0.1939	0.2395	0.7250	0.0572	0.2083	0.4568	0.4820	0.3364
LOG+BC-OLS	0.1833	0.2349	0.7470	0.0549	0.2388	0.5099	0.5238	0.3699
LOG+RiR	0.1854	0.2347	0.7492	0.0547	0.2400	0.4922	0.5274	0.3730
LOG+RoR	0.1877	0.2390	0.7451	0.0567	0.2118	0.4744	0.5190	0.3665
LOG+RT	0.1846	0.2344	0.7380	0.0547	0.2420	0.5000	0.4903	0.3445
LOG+MARS	0.1738	0.2217	0.7726	0.0486	0.3215	0.5687	0.5597	0.3985
LOG+LSSVM	0.1708	0.2197	0.7835	0.0479	0.3340	0.5797	0.5795	0.4163
LOG+ANN	0.1689	0.2188	0.7810	0.0476	0.3396	0.5845	0.5737	0.4135
OLS+RT	0.1779	0.2320	0.7660	0.0530	0.2572	0.5357	0.5554	0.3963
OLS+MARS	0.1713	0.2215	0.7740	0.0484	0.3231	0.5769	0.5707	0.4082
OLS+LSSVM	0.1695	0.2216	0.7882	0.0485	0.3223	0.5755	0.5933	0.4279
OLS+ANN	0.1747	0.2277	0.7730	0.0510	0.2844	0.5567	0.5706	0.4086

Table B.5: BANK5 model performance results

Technique	**MAE**	**RMSE**	**AUC**	**AOC**	R^2	r	ρ	τ
OLS	0.2085	0.2874	0.7180	0.0822	0.1197	0.3502	0.3032	0.2071
B-OLS	0.1783	0.3055	0.7120	0.0933	0.0933	0.3054	0.3112	0.2138
BR	0.2612	0.3019	0.7090	0.0909	0.1029	0.3209	0.3138	0.2151
BC-OLS	0.1824	0.3149	0.7100	0.0988	0.0815	0.2855	0.3139	0.2172
RiR	0.2086	0.2868	0.7200	0.0818	0.1231	0.3544	0.3045	0.2076
RoR	0.2087	0.2875	0.7180	0.0822	0.1189	0.3493	0.3030	0.2070
RT	0.2061	0.2885	0.7040	0.0829	0.1129	0.3390	0.3180	0.2482
MARS	0.2057	0.2856	0.7184	0.0811	0.1302	0.3615	0.3131	0.2251
LSSVM	0.2031	0.2812	0.7360	0.0787	0.1570	0.3964	0.3207	0.2190
ANN	0.2004	0.2860	0.7210	0.0815	0.1281	0.3619	0.2893	0.2000
LOG+OLS	0.2086	0.2876	0.7180	0.0824	0.1182	0.3479	0.3012	0.2060
LOG+B-OLS	0.1899	0.2964	0.7070	0.0875	0.0635	0.3225	0.2913	0.2000
LOG+BR	0.2875	0.3204	0.7070	0.1024	-0.0946	0.3346	0.2806	0.1918
LOG+BC-OLS	0.1863	0.3055	0.7120	0.0930	0.0963	0.3103	0.3050	0.2118
LOG+RiR	0.2062	0.2933	0.7128	0.0856	0.0831	0.3409	0.3150	0.2176
LOG+RoR	0.2060	0.2937	0.7118	0.0858	0.0806	0.3391	0.3162	0.2191
LOG+RT	0.2052	0.2890	0.6880	0.0832	0.1100	0.3348	0.3179	0.2219
LOG+MARS	0.2058	0.2934	0.6942	0.0857	0.0820	0.3285	0.2953	0.2121
LOG+LSSVM	0.2024	0.2887	0.7191	0.0829	0.1116	0.3652	0.3159	0.2190
LOG+ANN	0.2038	0.2854	0.7290	0.0811	0.1319	0.3689	0.3243	0.2216
OLS+RT	0.2066	0.2866	0.7190	0.0817	0.1244	0.3623	0.3067	0.2100
OLS+MARS	0.2068	0.2861	0.7237	0.0815	0.1271	0.3634	0.3081	0.2102
OLS+LSSVM	0.2087	0.2875	0.718	0.0822	0.1189	0.3493	0.3030	0.2070
OLS+ANN	0.2085	0.2874	0.7190	0.0822	0.1200	0.3498	0.3049	0.2086

Table B.6: BANK6 model performance results

References

[1] Basel Committee On Banking Supervision. **International Convergence of Capital Measurements and Capital Standards**. Technical report, Bank for International Settlements, 1988. 2, 3

[2] Basel Committee on Banking Supervision. **Basel II: International Convergence of Capital Measurement and Capital Standards: a Revised Framework**. Technical report, Bank for International Settlements, 2004. 2, 4, 5, 6, 7, 8, 10, 11, 12, 13

[3] Basel Committee on Banking Supervision. **Basel III: A global regulatory framework for more resilient banks and banking systems**. Technical report, Bank for International Settlements, 2010. 2, 5

[4] Basel Committee on Banking Supervision. **Basel III: International framework for liquidity risk measurement, standards and monitoring**. Technical report, Bank for International Settlements, 2010. 2, 5

[5] Basel Committee on Banking Supervision. **An Explanary Note on the Basel II IRB Risk Weight Functions**. Technical report, Bank for International Settlements, 2004. 5, 135, 136, 137, 138, 142, 144, 145

[6] Basel Committee on Banking Supervision. **Studies on the Validation of Internal Rating Systems, Working Paper No. 14**. Technical report, Bank for International Settlements, 2005. 5, 7, 8, 9, 10, 12, 15, 16, 17, 60, 61, 62, 78

[7] B. Baesens and T. Van Gestel. *Credit Risk Management: Basic Concepts*. Oxford University Press, USA, 2009. 6, 7, 8, 10, 13, 14

[8] N. Siddiqi. *Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring*. Wiley, 2005. 6

[9] Basel Committee on Banking Supervision. **QIS Frequently Asked Questions (as of 20 December 2002)**. Technical report, Bank for International Settlements, 2010. 6

[10] G. Moral and R. Garcia. **LGD Estimates in a Mortgage Portfolio**. *Banco de Espana, Estabilidad Financiera, No. 3*, 2002. 7

[11] K. Emery, S. Ou, J. Tennant, A. Matos, and R. Cantor. **Corporate Default and Recovery Rates**. Technical report, Moody's Global Credit Research, 2009. 7

[12] C. Needham and M. Verde. **Fitch Ratings Global Corporate Finance 2009 Transition and Default Study**. Technical report, Credit market research, Fitch Ratings, 2009. 7

[13] D. Vazza, D. Aurora, and N. Kraemer. **Annual Global Corporate Default Study And Rating Transitions**. Technical report, Global Fixed Income Research, Standard's and Poor, 2009. 7

[14] H. Almeida and T. Phillipon. **The Risk-Adjusted Cost of Financial Distress**. *Journal of Finance*, **6**:2557–2586, 2007. 8

[15] B. Brady, P. Chang, P. Miu, B. Ozdemir, and D. Schwartz. **Discount Rate for Workout Recoveries: An empirical study**. 2006. 8, 9

[16] I. Maclachlan. **Choosing the Discount Factor for Estimating Economic LGD**. 2004. 8, 9

[17] G. Moral and R. Garcia. **LGD Estimates in a Mortgage Portfolio**. *Banco de Espana,Estabilidad Financiera,No. 3*, 2002. 9

[18] T. Schuermann. **What Do We Know About Loss Given Default**. 2004. 10

[19] G. Bakshi, D. Madan, and F. Zhang. **Understanding the Role of Recovery in Default Risk Models: Empirical Comparisons and Implied Recovery Rates**. Technical report, FDIC Center for Financial Research, 2006. 10

[20] H. Unal, D. Madan, and L. Guntay. **Pricing the Risk of Recovery in Default with APR Violations**. *Journal of Banking and Finance*, **27(6)**:10011025, 2003. 10

[21] M.J. Rowan, P. Stump, E. De Bodard, and D. Staples. **Probability of Default and Loss Given Default Assessments**. Technical report, Moody's Corporate Finance, 2006. 10

[22] R. Merrit and R. Hunter. **Recovery Ratings: Exposing the Components of Credit Risk**. Technical report, Credit Policy, Fitch Ratings, 2005. 10

[23] S.B. Samson. **Corporate Ratings Criteria**. Technical report, Standard and Poor's, 2006. 10

[24] O.O. Maimon and L. Rokach. *Decompositional Methodology for Knowledge Discovery and Data Mining: Theory and Applications (Machine Perception and Artificial Intelligence)*. World Scientific Publishing Company, 2005. 12

[25] T. Bellotti and J. Crook. **Loss given default models incorporating macroeconomic variables for credit cards**. *International Journal of Forecasting*, **28**:171182, 2012. 12, 17, 18, 26, 27, 28, 29, 53, 63, 64, 66, 94

[26] E. Altman. **Default Recovery Rates and LGD in Credit Risk Modeling and Practice**. 2006. 13

[27] B. Baesens, Mues C. Setiono, R., and J. Vanthienen. **Using neural network rule extraction and decision tables for credit-risk evaluation**. *Management Science*, **49(3)**:312329, 2003. 14, 56

[28] B. Baesens, T. Van Gestel, M. Stepanova, D. Van den Poel, and J. Vanthienen. **Neural network survival analysis for personal loan data**. *Journal of Operation Research Society*, **59(9)**:10891098, 2005. 14

[29] D.J. Hand. **Modelling consumer credit risk**. *IMA Journal of Management Mathematics*, **12**:139155, 2001. 14

[30] M. Stepanova and L.C. Thomas. **Survival analysis methods for personal loan data**. *Operations Research*, **50(2)**:277289, 2002. 14

[31] L.C. Thomas, J. Ho, and W.T. Scherer. **Time will tell: Behavioural scoring and the dynamics of consumer risk assessment**. *IMA Journal of Management Mathematics*, **12**:89103, 2001. 14

[32] A.F. Atiya. **Bankruptcy prediction for credit risk using neural networks: A survey and new results**. *IEEE Transactions on Neural Networks*, **12(4)**:929935, 2001. 14

[33] L. Beccheti and J. Sierra. **Bankruptcy risk and productive efficiency in manufacturing firms**. *Journal of Banking and Finance*, **27**:20992120, 2002. 14

[34] D. Martens, B. Baesens, T. Van Gestel, and J. Vanthienen. **Comprehensible credit scoring models using rule extraction from support vector machines**. *European Journal of Operational Research*, **183**:1466–1476, 2007. 14

[35] M.L. Nasir, R.I. John, and S.C. Bennett. **Predicting corporate bankruptcy using modular neural networks**. In *Conference on Computational Intelligence for Financial Engineering*, 2000. 14

[36] T. Van Gestel, B. Baesens, J.A.K. Suykens, D. Van den Poel, D. Baestaens, and M. Willekens. **Bayesian kernel based classification for financial distress detection**. *European Journal of Operational Research*, **172**:9791003, 2006. 14

[37] M. Klopfer. *Gestion financire des collectivits locales*. Guides et mthodes, 2010. 14

[38] D. Laster. **Insurance company ratings**. Technical report, Sigma 4, Swiss Re, 2003. 14

[39] A. Estrella, S. Peristiani, and S. Park. *Credit ratings, methodologies, rationale, and default risk*, chapter Capital Ratios and Credit Ratings as Predictors of Bank Failures, page 233256. Risk Books, London, UK, 2002. 14

[40] A.E. Kocagil, A. Reyngold, R.M. Stein, and E. Ibarra. **Moodys RiskCalcTM Model for Privately-Held US Banks**. Technical report, Global credit research, Moodys Investors Service., 2002. 14

[41] A. Le Bras and D. Andrews. **Bank rating methodology**. Technical report, Fitch Ratings, 2003. 14

[42] S. Sarkar and R. Sriram. **Bayesian models for early warnings of bank failures**. *Management Science*, **47(10)**:14571475, 2001. 14

[43] T. Van Gestel, B. Baesens, P. Van Dijcke, J. Garcia, J. Suykens, and T. Alderweireld. **Linear and nonlinear credit scoring by combining logistic regression and Support Vector Machines**. *Journal of Credit Risk*, **1(4)**:32–60, 2005. 14

[44] D.C. Wheelock and P.W. Wilson. **Why do banks disappear? The determinants of US bank failures and acquisitions**. *Review of Economics and Statistics*, **82(1)**:127138, 2000. 14

[45] V.V. Acharya, S.T. Bharath, and A. Srinivasa. **Understanding the Recovery Rates on Defaulted Securities**. 2004. 15, 79

[46] E. Altman and V.M. Kishore. **Almost Everything You Wanted to Know About Recoveries on Defaulted Bonds**. *Financial Analysts Journal*, **52 (6)**:57–64, 1996. 15, 79

[47] E. Altman, A. Resti, and A. Sironi. **Analyzing and Explaining Default Recovery Rates**. 2001. 15, 79

[48] E. Altman, A. Resti, and A. Sironi, editors. *Recovery Risks: The Next Challenge in Credit Risk Management*. Recovery Books, 2005. 15, 79, 144

[49] M. Araten, M. Jacobs, and P. Varshney. **Measuring LGD on Commercial Loans: An 18-Year Internal Study**. *RMA Journal*, **86**:28–35, 2004. 15, 26, 79

[50] E. Asarnow and D. Edwards. **Measuring Loss on Defaulted Bank Loans: A 24-Year Study**. *Journal of Commercial Bank Lending*, **77**:11–23, 1995. 15, 26, 79

[51] L. Carty and D. Lieberman. **Defaulted Bank Loan Recoveries**. *Moody's Investors Service*, 1996. 15, 79

[52] L.V. Carty, D.T. Hamilton, S.C. Keanan, A. Moss, T. Mulvaney, T. Marshella, and M.G. Subhas. **Bankrupt Bank Loan Recoveries**. *Moody's Investors Service*, 1998. 15, 79

[53] R. Eales and E. Bosworth. **Severity of Loss in the Event of Default in Small Business and Large Consumer Loans**. *Journal of Lending and Credit Risk Management*, **80 (9)**:58–65, 1998. 15, 79

[54] J. Frye. **LGD in High Default Years**. *Federal Reserve Bank of Chicago*, 2003. 15, 79

[55] G.M. Gupton, D. Gates, and L.V. Carty. **Bank Loan Loss Given Default**. Technical report, Moody's Investors Service, 2000. 15, 79

[56] J. Frye. **Depressing Recoveries**. *Risk*, **13 (11)**:108–111, 2000. 15, 79

[57] D.T. Hamilton, P. Varma, S. Ou, and R. Cantor. **Default and Recovery Rates of Corporate Bond Issuers: A Statistical Review of Moody's Ratings Performance 1920-2002**. Technical report, Moody's Investors Service, 2003. 15, 79

[58] S. O'Shea, S. Bonelli, and R. Grossman. **Bank Loan and Bond Recovery Study: 1997-2000**. *Fitch Structured Finance*, 2001. 15

[59] J. Roche, W. Brennan, D. McGirt, and M. Verde. **Bank Loan Ratings in Bank Loans: Secondary Market and Portfolio Management**. 1998. 15

[60] B. Baesens, T. Van Gestel, S. Viaene, M. Stepanova, J. A. K. Suykens, and J. Vanthienen. **Benchmarking state of the art classification algorithms for credit scoring**. *Journal of the Operational Research Society*, **54**:627–635, 2003. 16

[61] J. A. Bastos. **Forecasting bank loans loss-given-default**. **34**:2510–2517, 2009. 17, 18, 26, 27, 28, 63, 65, 66, 76, 94

[62] G. Gupton. **Advancing Loss Given Default Prediction Models: How the Quiet Have Quickened**. *Economic Notes by Banca Monte dei Paschi di Siena SpA*, **34**:185230, 2005. 17, 18, 26, 27, 28, 29, 63, 64, 66, 76, 94

[63] J.A. Bastos. **Predicting bank loan recovery rates with neural networks**. Technical report, Technical University of Lisbon, 2010. 17, 18, 27, 63, 65, 66, 94

[64] R. Calabrese. **Estimating bank loans loss given default by generalized additive models**. Technical report, University College Dublin, 2012. 17, 18, 26, 27, 28, 63, 66, 76, 94

[65] S. G. Caselli and F. Querci. **The sensitivity of the loss given default rate to systematic risk: New empirical evidence on bank loans**. *Journal of Financial Services Research*, **34**:1–34, 2009. 17, 18, 26, 27, 28, 29, 53, 64, 65, 76, 94

[66] R. Chalupka and J. Kopecsni. **Modeling Bank Loan LGD of Corporate and SME Segments: A Case Study**. *Czech Journal of Economics and Finance*, **59**:360–382, 2009. 17, 18, 26, 27, 28, 29, 64, 65, 76, 94

[67] J. Dermine and C. Neto de Carvalho. **Bank Loan Losses-Given-Default, a Case Study**. *Journal of Banking and Finance*, 2005. 17, 18, 26, 27, 28, 29, 53, 64, 76, 94

[68] J. Grunert and M. Webera. **Recovery Rates of Bank Loans: Empirical Evidence for Germany**. Technical report, University of Mannheim, 2006. 17, 18, 26, 27, 28, 53, 64, 65, 94

[69] M. Bruche and C. Gonzlez-Aguado. **Recovery rates, default probabilities, and the credit cycle**. *Journal of Banking and Finance*, **34**:754–764, 2010. 17, 18, 27, 63, 66, 94

[70] G. Castermans, D. Martens, T. Van Gestel, B. Hamers, and B. Baesens. **An overview and framework for PD backtesting and benchmarking**. *Journal of the Operational Research Society*, pages 1–15, 2009. 18, 21, 61, 81, 88

[71] G. Christodoulakis and S. Satchell. *The analytics of risk model validation*. Elsevier, 2008. 18, 61

[72] B. Engelmann and R. Rauhmeier. *The Basel II Risk Parameters: Estimation, Validation, and Stress Testing*. Springer, 2006. 18, 61

[73] P. Brazdil and C. Soares. **Ranking learning algorithms: using IBL and meta-learning on accuracy and time results**. *Machine Learning*, **50**:251–277, 2003. 18, 94, 98, 100

[74] A. Kalousis and M. Hilario. **Model selection via meta-learning**. *International Journal on AI Tools*, **10**:4, 2001. 18, 94, 98, 100

[75] C. Kopf, C. Taylor, and J. Keller. **Meta-analysis: From data characterisation for meta-learning to meta-regression**. In *Proceedings of the PKDD Workshop on Data Mining, Decision Support, Meta-Learning and ILP*, 2000. 18, 94, 98, 101

[76] C. Linder and R. Studer. **AST: support for algorithm selection with a CBR approach**. In *Proceedings of the 16th International Conference on Machine Learning*, 1999. 18, 94, 98, 100

[77] D. Michie. *Machine Learning, Neural and Statistical Classification*, chapter 10. 1994. 18, 94, 98, 100, 102

[78] S. Ali and K. A. Smith. **Kernel width selection for SVM classification: A meta learning approach**. *International Journal of Data Warehousing and Data Mining*, **1**:78–97, 2005. 18, 94, 100, 102

[79] J. Demsar. **Statistical Comparison of Classifiers over Multiple Data Sets**. *Journal of Machine Learning Research*, **7**:1–30, 2006. 20, 22, 29, 51, 112, 122

[80] S. Garcia and F. Herrera. **An Extension onStatistical Comparisons of Classifiers over Multiple Data Setsfor all Pairwise Comparisons**. *Journal of Machine Learning Research*, **9**:2677–2694, 2008. 20, 22, 30, 51, 52, 112, 113, 122

[81] M. Friedman. **A comparison of alternative tests of significance for the problems of m rankings**. *Annals of Mathematical Statistics*, **11**:86–92, 1940. 21, 22, 29, 51, 112, 121

[82] G. Hommel. **A stagewise rejective multiple test procedure based on a modified Bonferroni test**. *Biometrika*, **75**:383–386, 1988. 21, 30, 51

[83] R. S. Witte and J. S. Witte. *Statistics*. Wiley, 2009. 21, 70

[84] A. R. Ansari and R. A. Bradley. **Rank-Sum Tests for Dispersions**. *The Annals of Mathematical Statistics*, **31**:1174–1189, 1960. 21, 71

[85] F. Wilcoxon. **Individual Comparisons by Ranking Methods**. *Biometrics Bulletin*, **1**:80–83, 1945. 21, 69

[86] J. Bi and K. P. Bennet. **Regression Error Characteristic Curves**. In *Twentieth International Conference on Machine Learning*, 2003. 21, 27, 44, 65

[87] N. Draper and H. Smith. *Applied Regression Analysis*. Wiley, 1998. 21, 31, 44

[88] G. Loterman, I. Brown, D. Martens, C. Mues, and B. Baesens. **Benchmarking regression algorithms for loss given default modeling**. *International Journal of Forecasting*, **28**:161–170, 2012. 21, 61, 63, 64, 66, 77

[89] T. Fawcett. **An introduction to ROC analysis**. *Pattern Recognition Letters*, **27**:861–874, 2006. 21, 43, 64

[90] J. R. Rice. **The Algorithm Selection Problem**. *Advances in Computers*, **15**:65–118, 1976. 22, 96

[91] Soares C. Prudncio, R.B.C and T.B. Ludermir. **Uncertainty Sampling-Based Active Selection of Datasetoids for Meta-learning**. In *21st international conference on artificial neural networks*, 2011. 22, 95, 101, 110

[92] C. Soares. **UCI++: Improved support for algorithm selection using datasetoids**. *Advances in Knowledge Discovery and Data Mining*, **5476**:499–506, 2009. 22, 95, 101, 110

[93] J. Alcal-Fdez, A. Fernandez, J. Luengo, J. Derrac, S. Garca, L. Snchez, and F. Herrera. **KEEL Data-Mining Software Tool: Data Set Repository, Integration of Algorithms and Experimental Analysis Framework**. *Journal of Multiple-Valued Logic and Soft Computing*, **17**:255–287, 2011. 22, 95, 109, 111

[94] S. Holm. **A Simple Sequentially Rejective Multiple Test Procedure**. *Scandinavian Journal of Statistics*, **6**:65–70, 1979. 22, 112, 122

[95] M. Gurtler and M. Hibbeln. **Improvements in loss given default forecasts for bank loans**. *Journal of Banking and Finance*, 2013. 26, 29

[96] C. Friedman and S. Sandow. **Ultimate recoveries**. *Risk*, **16**:6973, 2003. 26

[97] O. Renault and O. Scaillet. **On the Way to Recovery: A Nonparametric Bias Free Estimation of Recovery Rate Densities**. *Journal of Banking and Finance*, **28**:2915–2931, 2004. 26

[98] P. Cohen, J. Cohen, S.G. West, and L.S. Aiken. *Applied multiple regression/correlation analysis for the behavioral sciences*. Lawrence Erlbaum, 2002. 28, 45, 46

[99] P. Nemenyi. *Distribution-free multiple comparisons*. PhD thesis, Princeton University, 1963. 30

[100] A. E. Hoerl and R. W. Kennard. **Ridge Regression: Biased Estimation for Nonorthogonal Problems**. *Technometrics*, **12**:55–67, 1970. 32

[101] P.W. Holland and R.E. Welsch. **Robust Regression Using Iteratively Reweighted Least Squares**. *Communications in Statistics: Theory and Methods*, **6**:813 – 827, 1977. 32

[102] P.J. Huber. **Robust Estimation of a Location Parameter**. *Annals of Mathematical Statistics*, **35**:73–101, 1964. 32

[103] P.J. Huber and E.M. Ronchetti. *Robust statistics*. Wiley, 2009. 33

[104] G.M. Gupton and R.M. Stein. **LossCalc TM: Moody's Model for Predicting Loss Given Default**. Technical report, Rating methodology, Moodys, 2002. 33

[105] M. Smithson and J. Verkuilen. **A better lemon squeezer?Maximum-likelihood regression with beta-distributed dependent variables**. *Psychological Methods*, **11**:54–71, 2006. 34

[106] G.E.P. Box and D.R. Cox. **An Analysis of Transformations**. *Journal of Royal Statistics Society*, **26**:211–252, 1964. 35

[107] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen. *Classification and Regression Trees*. Chapman & Hall/CRC, 1984. 36, 105

[108] J. F. Friedman. **Multivariate Adaptive Regression Splines**. *The Annals of Statistics*, **19**:1–141, 1991. 37, 105

[109] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995. 38

[110] J.A.K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle. *Least Squares Support Vector Machines*. World Scientific Publishing Company, 2003. 38

[111] H. Wang and D. Hu. **Comparison of SVM and LS-SVM for Regression**. *International Conference on Neural Networks and Brain*, **1**:279–283, 2005. 38

[112] C.M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995. 39

[113] T. Van Gestel, D. Martens, D. Feremans, B. Baesens, J. Huysmans, and J. Vanthienen. **Forecasting and Analyzing Insurance Companies'Ratings**. *International Journal of Forecasting*, **23**:513–529, 2007. 40

[114] D.W. Hosmer and L. Stanley. *Applied Logistic Regression*. Wiley, 2nd edition edition, 2000. 41

[115] F. Hampel, R. Ronchetti, P.J. Rousseeuw, and W.A. Stahel. *Robust statistics : the approach based on influence functions*. Wiley, 1986. 49

[116] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining,Inference,and Prediction*. Springer Series in Statistics, 2001. 50, 105

[117] B. Baesens, S. Viaene, T. Van Gestel, J. A. K. Suykens, G. Dedene, B. De Moor, and J. Vanthienen. **An Empirical Assessment of Kernel Type Performance for Least Squares Support Vector Machine Classifiers**. *International Conference on Knowledge-Based Intelligent Engineering Systems and Allied Technologies*, **1**:313–316, 2000. 50

[118] T. Van Gestel, J.A.K. Suykens, B. Baesens, S. Viaene, J. Vanthienen, G. Dedene, B. De Moor, and J. Vandewalle. **Benchmarking Least Squares Support Vector Machine Classifiers**. *Machine Learning*, **54**:5–32, 2003. 50

[119] R. Freund and R. Littell. *SAS System for Regression*. Wiley, 2000. 51, 105

[120] R. Beran. **Simulated Power Functions**. *The Annals of Statistics*, **14**:151–173, 1986. 74, 82

[121] P. Hall and S.R. Wilson. **Two guidelines for bootstrap hypothesis testing**. *Biometrics*, **47**:757–762, 1991. 74, 82

[122] P.H. Westfall. *Re-sampling based multiple testing: examples & methods for p-Value adjustment*. Wiley, 1993. 74, 82

[123] K. Yuan. **Bootstrap Approach to inference and power analysis based on three test statistics for covariance structure models**. *British Journal of Mathematical and Statistical Psychology*, **56**:93110, 2003. 74, 82

[124] **Equal Credit Opportunity Act.** Technical report, United States Code, 1974. 77, 92

[125] N. J. D. Nagelkerke. **A Note on a General Definition of the Coefficient of Determination**. *Biometrica*, **3**:691–692, 1991. 79

[126] J. Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Routledge, 1988. 81, 82

[127] M. Svec. **PD backtest empirical study on credit retail portfolio**. 81

[128] B. Bade, D. Rosch, and H. Scheule. **Empirical performance of loss given default prediction models**. *Journal of Risk Model Validation*, **5**:2544, 2011. 88

[129] J. Huysmans, K. Dejaeger, C. Mues, J. Vanthienen, and B. Baesens. **An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models**. *Decision Support Systems*, **51**:141–154, 2011. 93

[130] D. G. Wolpert. **The Lack of A Priori Distinctions between Learning Algorithms**. *Neural Computation*, page 13411390., 1996. 93

[131] D. G. Wolpert. **Any Two Learning Algorithms Are (Almost) Exactly Identical**. Technical report, NASA Ames Research Center, 2001. 93

[132] J. Furnkranz and J. Petrak. **An evaluation of landmarking variants**. In *Proceedings of the ECML/PKDD Workshop on Integrating Aspects of Data Mining, Decision Support and Meta-Learning*, 2001. 94, 99, 102

[133] R. Leite and P. Brazdil. **Predicting relative performance of classifiers from samples**. In *22nd international conference on Machine learning*, 2005. 94, 99, 102

[134] C. Soares, J. Petrak, and P. Brazdil. *Progress in Artificial Intelligence*, chapter Sampling-Based Relative Landmarks: Systematically Test-Driving Algorithms before Choosing, pages 88–95. Lecture Notes in Computer Science, 2001. 94, 99, 102

[135] L. Rendell and H. Cho. **The effect of data character on empirical concept learning**. *Machine Learning*, **5**:267–298, 1990. 98

[136] D. Aha. **Generalizing from case studies: a case study**. In *Proceedings of the 9th International Conference on Machine Learning*, 1992. 98, 100

[137] H. Bensusan. **Odd Bites into Bananas Don't Make You Blind: Learning about Simplicity and Attribute Addition**. In *Proceedings of the ECML Workshop on Upgrading Learning to the Meta-level*, 1998. 99

[138] H. Bensusan, C. Giraud-Carrier, and C. Kennedy. **A Higher-order Approach to Meta-learning.** In *Proceedings of the ECML Workshop on Meta-learning: Building Automatic Advice Strategies for Model Selection and Method Combination*, 2000. 99

[139] Y. Peng, P. A. Flach, P. Brazdil, and C. Soares. **Improved Data Set Characterisation for Meta-Learning**. In *Proceedings of the Fith International Conference on Discovery Science*, 2002. 99

[140] H. Bensusan and C. Giraud-Carrier. **Discovering Task Neighbourhoods through Landmark Learning Performances**. In *Proceedings of the Fourth European Conference on Principles and Practice of Knowledge Discovery in Databases*, 2000. 99

[141] B. Pfahringer, H. Bensusan, and C. Giraud-Carrier. **Meta-learning by Landmarking Various Learning Algorithms**. In *Proceedings of the Seventeenth International Conference on Machine Learning*, 2000. 99, 101

[142] P. Kuba, P. Brazdil, C. Soares, and A. Woznica. **Exploiting sampling and meta-learning for parameter setting for support vector machines**. In *Proceedings of the Workshop Learning and Data Mining Associated*, 2002. 99

[143] C. Soares, P. Brazdil, and P. Kuba. **A meta-learning method to select the kernel width in support vector regression**. *Machine Learning*, **54**:195–209, 2004. 100

[144] J. Gama and P. Brazdil. **Characterization of classification algorithms**. In *Proceedings of the 7th Portugese Conference in AI*, 1995. 101

[145] G. Melli. **The datgen Dataset Generator**. 101

[146] D.A. Rachkovskij and E.M. Kussul. **DataGen: a generator of datasets for evaluation of classification algorithms**. *Pattern Recognition Letters*, **19**:537544, 1999. 101

[147] M. Reif, F. Shafait, and A. Dengel. **Dataset Generation for Meta-Learning**. Technical report, German Research Center for Artificial Intelligence, 2012. 101

[148] L.A. Garrow, T.D. Bodea, and M. Lee. **Generation of synthetic datasets for discrete choice analysis**. *Transportation*, **37**:183–202, 2010. 101

[149] P.D. Scott and E. Wilkins. **Evaluating data mining procedures: techniques for generating artificial data sets**. *Information and Software Technology*, **41**:579587, 1999. 101

[150] P. W. Frey and D. J. Slate. **Letter recognition using holland-style adaptive classifiers**. *Machine Learning*, **6**:161–182, 1991. 101

[151] K. A. Smith-Miles. **Cross-disciplinary perspectives on meta-learning for algorithm selection**. *ACM Computing Surveys*, **41**, 2008. 102

[152] D. C. Montgomery, E. A. Peck, and G. Geoffrey Vining. *Introduction to Linear Regression Analysis*. Wiley, 2012. 105

[153] A. C. Cameron, F. A.G. Windmeijer, H Gramajo, D. E. Cane, and C Khosla. **An R-squared measure of goodness of fit for some common nonlinear regression models**. *Journal of Econometrics*, **7**:329–342, 1997. 106

[154] B.S. Everitt. *Cambridge Dictionary of Statistics*. Cambridge University Press, 2002. 106

[155] H. Theil. *Economic forecasts and policy*. North-Holland Pub. Co., 1961. 107

[156] G. T. Knofczynski and D. Mundfrom. **Sample Sizes When Using Multiple Linear Regression for Prediction**. *Educational and Psychological Measurement*, **68**:431–442, 2008. 109

[157] M. Gordy. **A risk-factor model foundation for ratings-based bank capital rules**. *Journal of Financial Intermediation*, **12**:199–232, 2003. 138

[158] O. Vasicek. **Loan portfolio value.** *RISK*, pages 160–162, 2002. 140

[159] R.C. Merton. **On the pricing of corporate debt: The risk structure of interest rates.** *Journal of Finance*, **12**:449–470, 1974. 141

[160] H. Thomas and Z. Wang. **Interpreting the Internal Ratings-Based Capital Requirements**. *Journal of Banking Regulation*, **6**:274289, 2005. 141, 142

[161] A. De Servigny and O. Renault. **Measuring and managing credit risk**. *Risk*, **16**:90–94, 2003. 144

[162] G. Gupton and R. Stein. **LossCalc V2: Dynamic prediction**. Technical report, Rating methodology, Moodys, 2005. 144

[163] T. Schuermann. **Why were banks better off in the 2001 recession?** *Current Issues in Economics and Finance, Federal Reserve Bank of New York*, **10**, 2004. 144