DEPARTMENT OF ENVIRONMENT,
TECHNOLOGY AND TECHNOLOGY MANAGEMENT

Forecasting Loss Given Default Models: Impact of Account Characteristics and The Macroeconomic State

Ellen Tobback, David Martens, Tony Van Gestel & Bart Baesens

UNIVERSITY OF ANTWERP
Faculty of Applied Economics

EARNED EXCELLENCE
AACSB ACCREDITED
EST. 1916
THE BEST BUSINESS SCHOOLS IN THE WORLD

Stadscampus
Prinsstraat 13, B.226
BE-2000 Antwerpen
Tel. +32 (0)3 265 40 32
Fax +32 (0)3 265 47 99
http://www.ua.ac.be/tew

FACULTY OF APPLIED ECONOMICS

DEPARTMENT OF ENVIRONMENT,
TECHNOLOGY AND TECHNOLOGY MANAGEMENT

Forecasting Loss Given Default Models: Impact of Account Characteristics and The Macroeconomic State

Ellen Tobback, David Martens, Tony Van Gestel & Bart Baesens

RESEARCH PAPER 2012-019
SEPTEMBER 2012

University of Antwerp, City Campus, Prinsstraat 13, B-2000 Antwerp, Belgium
Research Administration – room B.226
phone: (32) 3 265 40 32
fax: (32) 3 265 47 99
e-mail: joeri.nys@ua.ac.be

The papers can be also found at our website:
www.ua.ac.be/tew (research > working papers) &
www.repec.org/ (Research papers in economics - REPEC)

D/2012/1169/019

Forecasting Loss Given Default Models : Impact of Account Characteristics and The Macroeconomic State

Ellen Tobback[a,*], David Martens[a,*], Tony Van Gestel[b,*], Bart Baesens[c,*]

[a] *Faculty of Applied Economics, University of Antwerp, Prinsstraat 13, 2000 Antwerp, Belgium*
[b] *Credit Risk Modelling, Group Risk Management, Dexia Group, Square Meeus 1, 1000 Brussels, Belgium*
[c] *Department of Decision Sciences & Information Management, K.U.Leuven, Naamsestraat 69,3000 Leuven, Belgium*

Abstract

Based on two datasets containing Loss Given Default (LGD) observations of home equity and corporate loans, we consider non-linear and non-parametric techniques to model and forecast LGD. These techniques include non-linear Support Vector Regression (SVR), a regression tree and a two-stage model combining a linear regression with SVR. We compare these models with an ordinary least squares linear regression. In addition, we incorporate several macroeconomic variables to estimate the influence of the economic state on loan losses. We investigate whether a Box-Cox transformation of the macroeconomic features improves the linear regression model. Due to the instable distributions, both out-of-time and out-of-sample setups are considered. The two-stage model outperforms the other techniques when forecasting out-of-time, while the non-parametric regression tree is the best performer when forecasting out-of-sample. The complete non-linear SVR reports poor prediction results, both in comprehensibility and accuracy. The incorporation of macroeconomic variables significantly improves the prediction performance of most of the models. These conclusions can help financial institutions when estimating LGD under the Internal Ratings Based Approach of the Basel Accords in order to estimate the downturn LGD needed to calculate the capital requirements.

Keywords: Loss Given Default, data mining, prediction, Basel III

1. Introduction

The recent credit crisis has emphasized the importance of the regulatory requirements. In response to the inadequacies in the previous accords, the Basel Committee on Banking Supervision has developed the Basel III Accord. The Basel Accords mainly cover the following four different risk types: credit risk, liquidity risk, market risk and operational risk. In this paper we focus on credit risk. The overall framework set out by the Basel Committee obliges financial institutions to hold a certain amount of regulatory capital to withstand the effects of a downturn period. The revised regulation demands for a better quality of capital reserves. However, the calculation of this reserve has not changed with respect to the previous accord. The minimum amount of capital that financial institutions are required to hold as a buffer depends on the risks they are exposed to. The Internal Ratings Based approach of the accord allows institutions to use internal risk assessments as inputs to the capital requirement calculations. Hence, banks can build internal credit risk models for the estimation of the three major risk parameters: Probability of Default (PD), Loss Given Default (LGD) and Exposure at Default (EAD). The LGD needed for the capital requirements is the downturn LGD. (BCBS, 2006) A fairly extensive body of literature has been written on modeling and forecasting PD. However, with respect to LGD, research is lagging behind. Nonetheless, Loss Given Default is of crucial importance when calculating credit default risk, as it is the percentage of the remaining outstanding balance that banks will not be able to recover. Previous studies (Bellotti and Crook, 2011; Bastos, 2009) report a relatively poor model fit. Hence, we face the challenge of improving the average model fit to provide an accurate LGD prediction model. Schuerman (2004) found that Loss

[*] Corresponding authors

Given Default varies with the business cycle and Bellotti and Crook (2011) concluded in their study that the incorporation of macroeconomic variables improves the prediction accuracy. Table 1 contains a selection of the research conducted on modeling LGD. Four out of the five studies have incorporated macroeconomic variables in their LGD models.

Regarding the modeling techniques, Bastos (2009) found that a non-parametric regression tree outperforms the parametric fractional regression models. Next, in a large-scale LGD benchmarking study, Loterman et al. (2011) have found that non-linear models, and in particular Support Vector Machines, significantly improve the prediction performance. On the contrary, Bellotti and Crook (2011) conclude that the OLS regression technique generates the best prediction model. In this paper we will therefore build on all those previous findings and add macroeconomic variables to non-linear and non-parametric models in order to try to enhance the goodness-of-fit of the LGD prediction models. These prediction performances are then compared to that of a Least Squares multilinear regression model. The regression techniques we consider are Support Vector Regression, a regression tree, Least Squares multilinear regression, Least Squares combined with a Box-Cox transformation applied to the macroeconomic features and a two-stage model combining a linear regression with SVR. Most of the previous studies have either used non-linear (or non-parametric) techniques without macroeconomic variables or linear techniques with macroeconomic variables. Few research has been conducted on the combination of non-linear modeling with macroeconomic variables. Another contribution of this paper lies in the fact that both our datasets cover the recent crisis period while previous research on LGD modeling with macroeconomic features has not used data beyond the year 2005. We consider losses on US corporate and home equity line of credit defaults. The corporate dataset covers 987 loans that defaulted between 1984 and 2011, while the home equity dataset covers 17.346 loans that defaulted between 2002 and 2008.

Incorporating macroeconomic features in the input space allows us to draw conclusions on the exact influence of the economic state on LGD and to estimate downturn LGD. By estimating downturn LGD using a downturn period as input to the different models, we investigate the sensitivity of the models to the macro economy. This is relevant as the LGD that should serve as capital should be a downturn LGD that occurs during periods with high default rates. Such periods are rare, but can be highly devastating. A good assesment of the downturn LGD with possible non-linear macroeconomic relations, is highly relevant for banks.

This paper is organized as follows: Section 2 gives an overview of the dataset and the variables considered in the feature selection. Next, Section 3 describes the regression techniques and the experimental set-up. Section 4 reports and discusses the results obtained, and finally, Section 5 concludes the paper.

2. Datasets description

This research is based on two real-life datasets obtained from different financial institutions. The datasets contain LGD observations of defaulted loans. Both datasets differ in size and loan portfolio type, which is why we will discuss them separately in the first two subsections. The third subsection handles the macroeconomic variables considered in the feature selection.

2.1. Home equity line of credit (HELOC)

The first dataset contains 17.346 monthly LGD observations of a revolving credit line of an American bank. A revolving credit line is a facility that allows borrowers to withdraw the amount of money needed from a previously specified total amount that is accessible over a specific period of time. Examples of revolving credit are credit cards or credit facilities used by corporates to provide their liquidity. In this case, however, the revolving credit line is a home equity line of credit (HELOC), meaning that the borrower's equity in his house is placed as collateral. The due payments depend on the utilization of the credit line. For each observed account, several account variables are reported. The observations cover a period of four year. The first observation date being 31/12/2002 and the last being 31/01/2008, the observations cover the period of the house bubble and burst, which resulted in the credit crisis. Figure 1 shows the approximate distribution of this dataset. Schuermann (2004) describes the recovery rate (1-LGD) as being either very high or very low, which results in an approximately 'bimodal'loss distribution. Note that the LGD distribution spikes around full loss and that it is far from normal (though it is not completely 'bimodal'either).

Table 1: Literature table

Authors	Type	Country	Period	Macroeconomic variables	Technique	Sample	Best peformance	Influence macroeconomy
Caselli, Gatti & Querci 2009	11.649 loans to households and SME's	Italy	1990-2004	Annual GDP growth rate, total annual number of employed people, annual household consumption, gross annual investments, total annual production, gross annual available income	Multivariate regression	out-of-sample	RMSE = 0.0732	annual GDP growth rate and EMP
Baston 2009	374 loans to SME's	Portugal	1995-2000	None	Log-log, logistic model, regression tree	out-of-sample (oos) + out-of-time (oot)	best oos RMSE = 0.396 (regression tree), best oot RMSE = 0.365 (logistic)	Not considered
Bellotti, Crook 2011	55.000 credit cards accounts	UK	1999-2005	UK retail banks' base interest rates, UK unemployment level, UK earnings index	OLS, decision tree, Tobit, Least absolute value regression, transformations: fractional logit and beta transformation	out-of-time	best MSE = 0.151 (OLS)	Unemployment level,income,interest rates
Dermine and Neto de Carvalho 2005	374 loans to SME's	Portugal	1995-2000	Annual GDP rate of growth, interest rate on the loan, frequency of default in industry sector	Logistic, log-log fractional regression	no prediction: empirical	Pseudo R^2 = 0.13	
Altman et al. 2005	1000 defaulted US bonds	US	1982-2001	Annual GDP growth rate, annual return on S&P 500 stock index, default rates	Multivariate regression	out-of-sample	No influence from macroeconomic variables	

The reported account variables are displayed in Table 2.

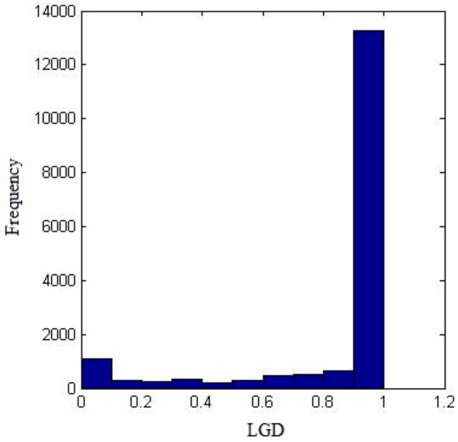
2.2. Corporate loans

The second dataset contains 986 yearly loss given default observations of a corporate loans portfolio. For each observed account, the credit rating, seniority and sector it belongs to are reported as well. The credit rating follows the external ratings from A- to CCC. For the seniority there are four possibilities, ranked from high to low: senior, senior unsecured, subordinated, junior subordinated. Seniority is expected to have a significant influence on the losses, as seniority refers to the order of debt repayment in case of bankruptcy. Senior loans should report lower losses than subordinated loans. Within the dataset, observations were divided into six different sectors: finance, industrial, insurance, public-utility, real estate finance, transportation and other non-bank.

The observations cover a period of 27 years, starting in 1984 and ending in 2011. Given that these are yearly observations, the selected values of the macroeconomic input variables are those that occurred on the first of January.

Figure 2 shows the approximate distribution of this dataset. Note that the distribution is different from that of the first dataset and far from ‘bimodal’.

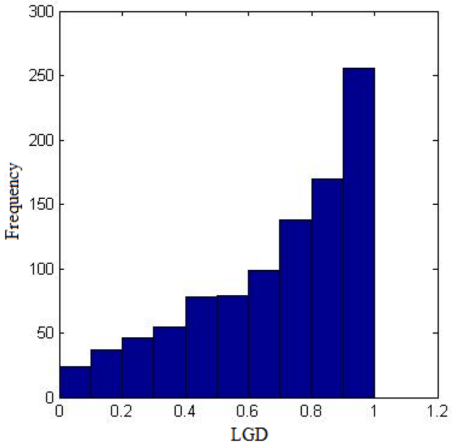
Figure 1: LGD distribution for the HELOC dataset.



2.3. Macroeconomic variables

Schuermann (2004) found that Loss Given Default varies with the business cycle and that losses tend to be higher during recessions. Therefore we can expect that LGD not only depends on the client and account characteristics but also on the state of the economy. Incorporating macroeconomic variables has thus the potential to improve the performance of LGD prediction models. We consider the following macroeconomic factors as candidate explanatory variables: unemployment rate, interest rates, exchange rates, GDP, equity prices, disposable personal income, inflation, confidence level and house prices.

Figure 2: LGD distribution for the Corporate dataset.



Whereas downturn can be understood as macroeconomic downturn, for banks and shareholders, the high impact comes from high default rates combined with high LGDs. This determines the capital need. Some market participants argue that during periods of high default rates, rather than macro-economic downturn, LGD levels rise as the vulture market gets saturated. Therefore, we have added default rates on investment and speculative bonds to candidate set of input variables. For each macroeconomic variable, several variants are included. Van Gestel et al.(2007) have shown that these variants have the potential to improve the performance as they provide relevant information related to previous periods. We define the following inputs:

1) 5 year moving average (av) :

$$\frac{x_t + x_{t-1} + x_{t-2} + \ldots + x_{t-4}}{5}$$

2) Growth rate (gr) :

$$\frac{x_t - x_{t-1}}{x_{t-1}}$$

3) Relative trend (rtr) (long term growth rate):

$$x_{tr} = \frac{x_t - x_{t-4}}{4 \times x_{t-4}}$$

4) Absolute trend (atr):

$$x_{tr} = \frac{x_t - x_{t-4}}{4}$$

5) Level (x_0) :

$$x_t$$

For both datasets *t, t-1, t-2, t-3, t-4* span a period of four years, except for the growth rate in the case of the home equity credit line. For this dataset monthly growth rates are used, as the default observations are reported monthly. This allows for responses to sudden shocks to be incorporated in the model. For some macroeconomic factors, the level and/or absolute trend is not considered as explanatory variable. This is the case for GDP, equity prices (stock index), disposable personal income and the S&P 500 house index. This is justified by the fact that these factors are non-stationary.[1] Incorporating these variables in their non-stationary form would result in erroneous conclusions about their effect on LGD. Regarding default rates we consider only the level and moving average.

3. Regression techniques and methodology

3.1. Regression techniques

We consider non-linear regression techniques and data transformations to model LGD. The performance of these models is compared to that of a simple multiple linear regression model. We consider a Box-Cox transformation on the macroeconomic variables, Support Vector Regression (SVR), a non-parametric regression tree and a two-stage model that combines the linear regression with a SVR regression on the residuals.

3.1.1. Box-Cox transformation

In a linear model, the explanatory variables are expected to influence the dependent variable in a linear way. This means that the impact of a two percentage increase in GDP on the LGD will have the same absolute magnitude as a two percentage decrease in GDP, which is very unlikely. Therefore, we consider an intrinsically linear model by applying non-linear transformations of the macroeconomic variables before fitting a linear model. (Box & Cox, 1964; Van Gestel & al, 2007). The non-linear transformation is a Box-Cox power transformation. The Box-Cox transformation 'family' has the following form (Box & Cox, 1964). Setting λ to a proper value is discussed in Section 3.2.3.

$$y_i^{(\lambda)} = \begin{cases} (y_i^\lambda - 1)/\lambda; & \lambda \neq 0 \\ \log(y_i); & \lambda = 0 \end{cases}$$

3.1.2. Regression tree

The major advantage of using a regression tree is that the interpretation is quite logical and simple. We want to predict the response LGD given a number of inputs (account variables and macroeconomic variables). When setting up a tree, at each node a test will be applied to one of the variables. The response to this question will decide whether the right or left branch of the tree should be further followed. At the final node the predicted value equals the average value of all observations that ended up here during the training. A regression tree is thus a non-parametric regression technique that is based on the average target variable of the observations that measure up to certain conditions. The regression trees are created using the CART algorithm. (Breiman et al.,1984).

3.1.3. Non-linear Support Vector Regression

A more advanced, state-of-the-art technique under the classifiers are Support Vector Machines. With this technique the input vectors are mapped into a higher dimensional, kernel induced features space. In this paper, we use Support Vector Regression, which is the same concept as Support Vector Machines but applied to a regression instead of a classification. When building the SVR model we apply a RBF kernel function with parameters determined through a grid search procedure (Van Gestel et al., 2004). SVR can capture complex non-linear relationships within the data and is therefore expected to have a better prediction performance compared to the linear model we consider in this paper. However, SVR is less comprehensible as this is a 'black box'model. To understand the inside of the model, one could use rule extraction (Martens et al, 2007). However, this lies outside the scope of this paper. We consider

[1] e.g. the S&P 500 price index was 164.93 on 01/01/1984. Twenty-eight years later,on 01/01/2012, the index stood at 1257.6.

Table 2: Account variables

Variable	Description
LTV	Loan to Value
Age	The age of the account
PayOff	Principle balance augmented with unpaid interests and fees
Equifax	Equifax credit score. It indicates the riskiness of the loan.The score lies between 300 and 850, with a median value of 723. The higher the credit score of an individual, the higher the creditworthiness. A client with a credit score below 600 can be asked to pay a higher interest rate on the loan.
Utilization	Percentage of the credit line withdrawn by the client
PD	Probability of default

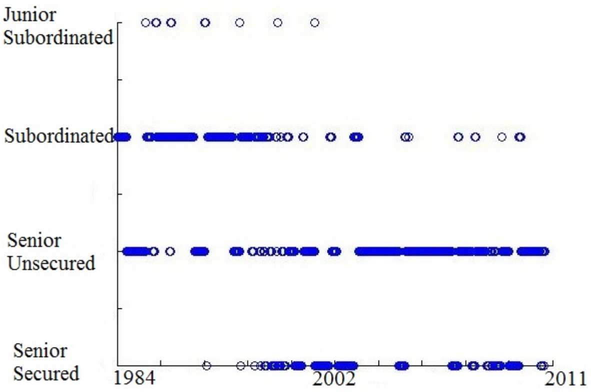
both a complete SVR and a combination between the linear technique and SVR. The latter approach first estimates a linear relationship between LGD and the features. In a second step SVR is used to capture non-linearities in the error terms. This approach thus combines the comprehensiveness of the linear technique with the expected accuracy of SVR, similarly to (Van Gestel et al., 2007).

3.2. Experimental set-up

3.2.1. Dataset preprocessing

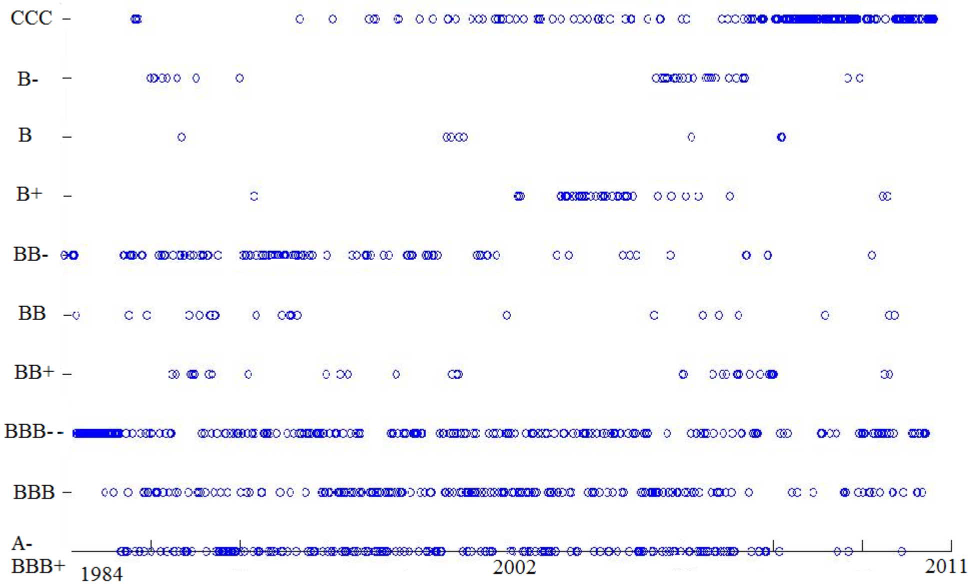
Both datasets are divided into two sets: two thirds training/validation set and one third test set. The former is used to train the system and estimate the model parameters, the latter is used for an out-of-sample forecast. The training/validation set is further divided into an actual training set and a validation set used for feature selection. In both cases the test set is out-of-time, as is the case when forecasts are made by financial institutions to estimate their future losses. Therefore the model evaluation based on the test set is a valid measurement of the prediction performance. However, as shown in Table 7 and discussed later, the out-of-time forecast reports a negative R^2 for the corporate dataset. Investigation of the distributions of the account variables for both the training and test set shows that there is a clear instability in the distribution of seniority and rating through time. Therefore we consider an out-of-sample though in-time forecast. Training, test and validation samples are chosen through stratification over these two variables. Figures 3 and 4 show the distribution of respectively seniority and rating over time in the form of a scatter plot. The division between the training and validation set on the one hand and the test set on the other hand is located at year 2002. This unequal distribution, which is the most obvious for junior subordinated in figure 3 and the rating CCC in figure 4, is the reason for the distortion in the prediction performance. Determining the training, test and validation samples through stratification over seniority and rating solves this problem (see Table 7).

Figure 3: Seniority



The corporate loans portfolio contains two ordinal independent variables, rating and seniority, and one nominal independent variable, sector. The nominal variable is dummy-encoded, rating and seniority are thermo-encoded. Thermometer (unary) encoding allows the inputs of a variable to be treated as ranked values. The rating score is a 15 notch external rating scale ranging from A- to CCC.

Figure 4: Rating



In our analysis, the last five notches are grouped to one, since only a few observations have a credit rating worse than CCC. This leaves us with 11 possible credit rating scores. However, since there are no observations with a rating score of A- in the training set when forecasting out-of-time, the first and second possible rating are grouped to one.

3.2.2. Outliers removal

Outliers in the explanatory variables can create a distortion in the weights assigned to these features. To avoid this possible distortion, outliers above or below the 3s borders are set equal to this cap or floor (Van Gestel et al., 2005). The boundaries are defined using the winsorised mean procedure where the maximum, respectively minimum allowed values are defined by $\max = \text{median}(x) + 3 \times s$ and $\min = \text{median}(x) - 3 \times s$ with $s = \frac{IQR(x)}{2 \times 0.6745}$ and IQR the interquartile range. Outliers are detected and reduced to the borders for the account variables age, payoff and loan-to-value.

3.2.3. Parameter settings

Both the Box-Cox and the SVR approach involve parameters that need to be set before implementing the technique. For SVR, these are the sigma (σ) and cost (C) parameter, while the Box-Cox approach requires a predefined λ coefficient. σ^2 and C are found conducting a grid search procedure using 5-fold cross-validation. For the out-of-time procedure, the grid search is performed out-of-time as well. For the out-of-sample procedure, the training set division into subsets is carried out through stratification. The limits for the regularization (C) and kernel (σ) hyper-parameters are set at $[\frac{0.01}{m}, \frac{1000}{m}]$ and $[0.5\sqrt{n}, 500\sqrt{n}]$ respectively, where m equals the number of input features and n the number of instances (Van Gestel et al., 2003). The first dataset contains approximately 7.200 training observations. For computational reasons we have selected a random and stratified[2] sample of 2.000 observations as input to the grid search.
The Box-Cox transformation requires a predefined λ coefficient for each variable that needs to be transformed. To define this parameter we rely on the built-in procedure in the used software.[3] Both the training and test set are transformed using the optimal λ parameter of the training set.

3.2.4. Feature selection

Input selection is used to eliminate those variables that are redundant to the estimation.

[2] Stratification is performed over Probability of Default, credit score and time

[3] Matlab's financial toolbox

A forward selection procedure based on the MSE value is applied in building the models. Only those variables that significantly improve the MSE value when added to the model are retained. For SVR, the first best variable to add is searched using the default values for the hyper-parameters. Next, a grid search is performed to determine the optimal values for the hyper-parameters with this variable as input. Based on the updated parameters the search is continued, and every five iterations the parameters are adjusted through a grid search procedure. The regression tree model is pruned up to the optimal level, based on the MSE value of the validation set.

3.2.5. Model evaluation

The model evaluation is based on five performance metrics: mean squared error (MSE), mean error (ME), mean absolute error (MAE), R-squared (R^2) and Spearman's rho. Both the MSE and the MAE give information about the expected difference between the actual future LGD observation and the predicted value (Draper et al., 1998). MSE differs from MAE in that it gives more weight to larger deviations from the actual observed LGD value. The mean error is a representation of the bias towards either an overestimation or underestimation of the predicted target variable. Preferably this value approximates zero. The R^2 value reports the proportion of the variability in the test sample that is explained by the model. The higher R^2, the better the model is able to capture this variability (Draper et al., 1998). Finally, Spearman's ρ is a non-parametric measure of correlation between the predicted and observed value. (Cohen et al., 2002)

4. Results and discussion

4.1. Feature selection

Tables 3, 4 and 5 show the variables added to the models of respectively the first (out-of-time and out-of-sample) and second (out-of-sample) dataset. The features selected for the out-of-time forecasts on the corporate dataset are not displayed in this paper, due to the fact that the forecasts report a negative R^2 and are thus not optimal. Many account features are selected to be incorporated into the models. Investigating the influence of these characteristics shows that the relation with LGD is as expected. A higher loan-to-value, utilization rate and payoff result in a higher LGD value. The less senior the loan and/or the lower the rating, the higher LGD. A loan from an industrial sector has a lower LGD, while public utility loans have higher LGDs than the other sectors.

4.2. Prediction performance

The values of the evaluation metrics of the models are displayed in Tables 6 and 7.

The out-of-time forecasts of the corporate dataset lead to a negative R^2 for each of the models. Spearman's ρ appears to be positive, though small. Due to the fact that the linear model reports a negative value for R^2, no Box-Cox transformation is performed on this model. Investigation of the input features shows a clear instability in the distribution of the account features seniority and rating as was discussed earlier and displayed in Figures 3 and 4. Hence, we cannot conclude that an out-of-time forecast is not possible. The out-of-time prediction performances of the HELOC dataset prove the fact that is possible to predict LGD out-of-time (though with a certain error). These performances evoke an important conclusion. While the test set of the HELOC dataset covers a crisis period (the 2007 credit crunch) resulting in a higher mean LGD and a different distribution of the macroeconomic features compared to the training and validation set, the models are still able to capture most of the variation in LGD and, moreover, report RMSE values that are competitive or even better than those reported in previous studies. This conclusion is important for financial institutions. When forecasting LGD, an equal distribution of the account features between the training set and the prediction portfolio is more important than an equal distribution of the macroeconomic features, which is impossible during a crisis period. It would be wrong to conclude that this means the macroeconomy has no influence and the next section proves otherwise. Though probably, LGD is mainly dependent on the first four out of the five C's of credit, which are character, collateral, capital and cash flows and to a smaller extent on the fifth C, conditions. Further research should be conducted on this topic.

Comparison of the out-of-sample and out-of-time forecasts of the HELOC dataset shows that in-time modeling is able to capture a larger share of the variation in the dataset than out-of-time modeling (higher R^2 and ρ). These results are in line with the expectations, since in-time forecasting includes future information on the macroeconomic state in the

Table 3: Input RCL out-of-time

Variable	Linear	Transformation	Regression tree	Linear + SVR	SVR
	I/O	I/O	I/O	SV	SV
Age	0	0	0	0	0
LTV	I	I	0	0	I
Payoff	0	0	I	0	0
Utilization	I	I	0	0	I
PD	0	0	0	0	0
Equifax	0	0	0	0	I
Unemployment rate	0	0	0	gr,rtr	0
Real GDP	0	rtr	0	rtr,gr	0
Inflation rate	0	0	0	av gr	0
Fed Fund rate	0	0	0	0	rtr
Spread	0	0	0	0	0
S&P500 index growth	0	0	0	x_0,rtr	0
US TWV	0	0	0	av gr	0
Income (DPI)	gr,rtr	0	0	rtr	0
S&P house index	0	0	0	0	0
defaultspeculative	0	0	0	0	0
defaultallbonds	0	0	0	0	0

Table 4: Input RCL out-of-sample

Variable	Linear	Transformation	Regression tree	Linear + SVR	SVR
	I/O	I/O	I/O	SV	SV
Age	0	0	0	I	0
LTV	I	I	I	0	I
Payoff	I	I	I	0	I
Utilization	I	I	0	0	I
PD	0	0	0	0	0
Equifax	0	0	0	I	I
Unemployment rate	0	gr,rtr	0	gr,rtr	0
Real GDP	0	0	0	0	0
Inflation rate	x_0	0	0	av gr	0
Fed Fund rate	0	0	0	0	0
Spread	0	0	0	0	0
S&P500 index growth	0	av gr	0	0	0
US TWV	0	x_0	0	0	0
Income (DPI)	gr	0	0	0	0
S&P house index	gr	0	gr	0	0
defaultspeculative	0	0	0	0	0
defaultallbonds	0	0	0	0	0

training set. The MSE and MAE values are higher for the out-of-sample forecast compared to the out-of-time prediction. This does not mean that out-of-sample forecasting is less accurate. The training and test sets of the two forecasting techniques are different and therefore the MSE and MAE statistics cannot be compared. Although in-time forecasting reports a higher sum of squared errors (which can be deducted when looking at the MSE values), it also reports a higher total sum of squares.

As discussed earlier we have considered several techniques to model LGD. Tables 6 and 7 report the prediction performances for each of the models. When forecasting out-of-time the two-stage model combining linear and SVR appears to be the best predictor. SVR on its own results in poor and for some statistics even the worst results. This could mean that SVR with a RBF kernel ignores existing

Table 5: Input Corporate out-of-sample

Variable	Linear	Transformation	Regression tree	Linear + SVR	SVR
	I/O	I/O	I/O	SV	SV
Seniority	3	3	2	1	2
Rating	6	4	2	1	4
Sector	0	0	2	1	1
Unemployment rate	0	0	0	0	0
Real GDP	0	0	gr	0	0
Inflation rate	av	0	0	av	rtr
Fed Fund rate	x_0,atr	0	0	0	0
Spread	av	0	av	0	0
S&P500 index	0	0	0	0	0
US TWV	0	0	atr	gr,rtr	atr,av
Consumer confidence	x_0	x_0	0	0	x_0,av
defaultspeculative	0	0	0	av	av
defaultallbonds	0	0	0	0	av

Table 6: Prediction performance results of the out-of-time and out-of-sample HELOC dataset for the linear model, transformed linear model, regression tree, SVR model and linear model with SV terms

Out-of-time

Technique	MSE	ME	MEA	R^2	ρ
Linear	0.0490	0.0395	0.1550	0.0671	0.1601
Box-Cox transformation	0.0515	0.0667	0.1731	0.0195	0.1750
Regression tree	0.0481	0.0763	0.1745	0.0850	0.2219
SVR	0.0525	0.0618	0.1763	0.0012	-0.0338
Linear + SVR	0.0473	0.0014	0.1338	0.0995	0.1745

Out-of-sample

Technique	MSE	ME	MEA	R^2	ρ
Linear	0.0706	-0.0059	0.1811	0.1626	0.2457
Box-Cox transformation	0.0755	0.0601	0.2158	0.1044	0.1928
Regression tree	0.0652	-0.0057	0.1684	0.2275	0.2745
SVR	0.0806	-0.0837	0.1420	0.0444	0.1461
Linear + SVR	0.0719	0.0139	0.1933	0.1472	0.1913

linearities in the relationship between LGD and the input space, possibly overfitting the data. The performance of the two stage model confirms that this relationship has both linear and non-linear components. This conclusion is conform with the expectations. An increase from 1% to 2% inflation is expected to have a different effect on the economy than an increase from 2% to 4%, though both increase with 100%. A linear model with this percentage as input would give equal weight to both. Fitting a SVR on the residuals accounts for these non-linearities.

When predicting in-time the regression tree reports the best results. This good performance is consistent with Bastos' (2010) findings on the performance of the regression tree when modeling out-of-sample LGD on a portfolio of loans to SMEs.

Contrary to expectations, though consistent with Bellotti and Crook's (2011) findings, least squares linear regression does not underperform when forecasting LGD. As the results in table 6 and 7 show, for two out of the four forecasting samples linear regression is the second best modeling technique. Using a Box-Cox transformation does not improve

Table 7: Prediction performance results of the out-of-time and out-of-sample corporate dataset for the linear model, transformed linear model, regression tree, SVR model and linear model with SV terms

Out-of-time					
Technique	MSE	ME	MEA	R^2	ρ
Linear	0.0985	-0.1377	0.2472	-0.2522	0.0975
Regression tree	0.0863	-0.0023	0.2382	-0.0968	0.0641
SVM	0.0927	-0.0888	0.239	-0.1777	0.0588
Linear + SVM	0.0849	-0.0088	0.2379	-0.0793	0.0746
Out-of-sample					
Technique	MSE	ME	MEA	R^2	ρ
Linear	0.0561	-0.0258	0.1836	0.0946	0.4015
Box-Cox transformation	0.0564	-0.0279	0.1849	0.090	0.3748
Regression tree	0.0534	-0.0415	0.1776	0.1389	0.4205
SVR	0.0592	-0.0324	0.1928	0.0449	0.2694
Linear + SVR	0.0572	-0.0415	0.1833	0.0776	0.4015

the results.

4.3. Macroeconomic influence and downturn LGD

Under the Basel accords banks are not allowed to set their LGD estimates for each portfolio lower than the long-run default-weighted average (BCBS, 2004). Table 8 represents the default-weighted average and unweighted average for both datasets. It appears that the default-weighted average is larger than the unweighted average, confirming the fact that an economic downturn period (which is thus accompanied by a higher number of defaults) results in a higher LGD for the portfolio. The downturn effect on LGD is calculated by subtracting the unweighted from the weighted average. The downturn effect is larger for the corporate dataset than for the home equity dataset. This could be caused by a possible higher exposure of corporate firms to counterparty credit risk. We compare the default-weighted average with LGD estimated by the different models during a recession. To represent the recession, actual observations of the macroeconomic features of December 2007 are used as input for the home equity line dataset. For the corporate dataset the observations of the macroeconomic features of January 2009 are chosen as they represent a more severe downturn period for the variables considered than the observations of January 2008. Choosing a different period means that the downturn values cannot be compared across datasets. However, both datasets have different properties which renders a comparison unreliable even when the same downturn period is considered. Therefore, we do not compare both datasets and choose two separate periods that best represent downturn movements. A selection of 10 combinations of account features is chosen. These combinations are chosen among those that are part of the out-of-time test set in order to use types of accounts on which defaults have actually occurred. The downturn LGD values are displayed in Table 9. The values of models that are not influenced by the macro economy are accompanied by an asterisk. These values are the same whatever the economic situation at the moment of default and are therefore not real downturn values. For the out-of-time HELOC and out-of-sample corporate models, the two stage model appears to be the most responsive to a downturn period. The regression tree model is not influenced by the macro economy. For the out-of-sample HELOC dataset, SVR and the two-stage regression technique report the highest mean LGD values, however, there are no macroeconomic variables incorporated in these two models.

The results suggest that there is no non-linear impact of the macro economy on loss given default. The linear model responds severely to the downturn conditions. However, the estimated model places no boundaries on the value of the dependent variable. It is therefore intuitive that the linear model responds strongly to stressed values.

Table 8: Default-weighted and unweighted average LGD and the difference between them (downturn effect) for the corporate and HELOC datasets

Dataset	Default-weighted Average	Unweighted Average	Downturn
Corporate	0.6903	0.6464	0.0439
Home equity line	0.8622	0.8540	0.0082

Table 9: Estimation of out-of-time and/or out-of-sample LGD in a downturn period for the corporate and HELOC datasets, using the linear model, transformed linear model, regression tree, SVR model and linear model with SV terms

HELOC					
Prediction	Linear	Transformation	SVR	Linear + SVR	Regression tree
Mean LGD (out-of-time)	0.8892	0.8816	0.8629	0.9350	0.84963*
Mean LGD (out-of-sample)	0.9483	0.8635	0.9703*	0.9483 + 0.0616*	0.9332
Corporate					
Mean LGD (out-of-sample)	0.7983	0.7138	0.7331	0.814	0.5205

Table 10: The effects of macroeconomic features included in the linear regression models (least squares and Box-Cox transformed) for the HELOC dataset

Out-of-time prediction		
Features	Linear regression	Box Cox transformation
Constant	−	+
Income rtr	+	0
Income growth	−	0
GDP rtr	0	+
Out-of-sample prediction		
Features	Linear regression	Box-Cox transformation
Constant	+	+
S&P house index growth	−	0
Inflation rate	−	0
Income growth	−	0
Unemployment rtr	0	−
S&P 500 av gr	0	−
Unemployment gr	0	−
US TWV	0	−

It is clear that the macroeconomic state influences the magnitude of the LGDs. The question remains which macroeconomic variables influence these losses and how. Tables 3 to 5 show which macroeconomic variables are incorporated in the models. The variables incorporated differ across the models and each macroeconomic feature is included in at least one of the models.

Figures 5, 6 and 7 show the regression trees of the corporate and HELOC datasets. The influence of the relative trend of GDP as represented in Figure 5 and the influence of the growth in housing prices as represented in Figure 7 is quite clear: the higher the relative GDP trend, the higher the

Table 11: The effects of macroeconomic features included in the linear regression models (least squares and Box-Cox transformed) for the corporate dataset

Features	Linear regression	Box-Cox transformation
Constant	+	+
Consumer Confidence	+	+
Inflation av	−	0
Fed Funds rate	+	0

LGD and the higher the growth in housing prices, the lower LGD. The first relationship is logical and confirms what happened during the credit crisis. The second relationship may seem less logical. However, previous studies have already confirmed the opposite relationship between GDP and credit risk (see e.g. Bonfim, 2009). During a period of high GDP growth, banks are willing to issue loans to more risky borrowers against a high return. Therefore, economic growth is accompanied by an accumulation of risks which result in greater losses once the growth starts slowing down. Moreover, when interpreting trend and growth variables, caution is required. The growth in, for instance, GDP is the highest when the economy starts to recover from a recession. The top of the business cycle, when the GDP level is maximal, is most of the time accompanied by a low growth percentage of GDP. The same relationship regarding GDP is found in the Box-Cox model for the out-of-time HELOC prediction. The signs of the macroeconomic variables for the linear models of the out-of-time and out-of-sample LGD prediction for the HELOC dataset are displayed in Table 10. The sign of the relative trend of income appears to be positive, though this compensates for the negative value of the constant. The negative sign of the income growth shows the real relationship between income and LGD: the higher the growth in disposable income per capita, the lower the LGD. Counter-intuitive signs are reported for the relative trend of unemployment, unemployment growth and the moving average of the inflation rate. For these three variables it appears that the higher the value, the lower LGD. Table 11 shows the same negative relationship between the moving average of the inflation rate and LGD. Next, the higher the federal funds rate and the higher the level of consumer confidence, the higher the losses. This positive influence of the confidence level can possibly be explained by the relation between economic growth and increasing risk. A higher Federal Funds rate makes it more expensive for borrowers to repay their loans. This can decrease their ability to pay off already defaulted loans.

Support vector regression creates a ‘black box’model as was mentioned earlier. This makes the models difficult to interpret and reduces the ability to estimate the influence of the economy on LGD. The out-of-time SVR prediction of the HELOC dataset, however, contains only one macroeconomic variable, the relative trend of the Federal Funds rate. We have estimated the LGD using one particular set of account variables combined with once a negative relative trend and once a positive relative trend of the Fed rate.It appears that a negative relative trend results in a higher LGD compared to a positive trend for the same account variables. The other SV models include more than one macroeconomic variable, making a conclusion on their influence ambiguous. Rule extraction should be used for this purpose. As stated earlier, this lies beyond the scope of the paper and constitutes an interesting issue for future research.

5. Conclusion

The results show that the complete non-linear model, support vector regression, is not fit for LGD forecasting based on the two datasets used in this research, because of its black-box nature and bad predictive behavior. The best performances are reported for the two-stage model combining a linear regression with a support vector regression on the error terms, when forecasting out-of-time. Forecasting out-of-sample is best conducted using a regression tree. The least squares linear regression does not underperform and can be used when a larger comprehensibility is required. Though, for

Figure 5: Regression tree Corporate out-of-sample

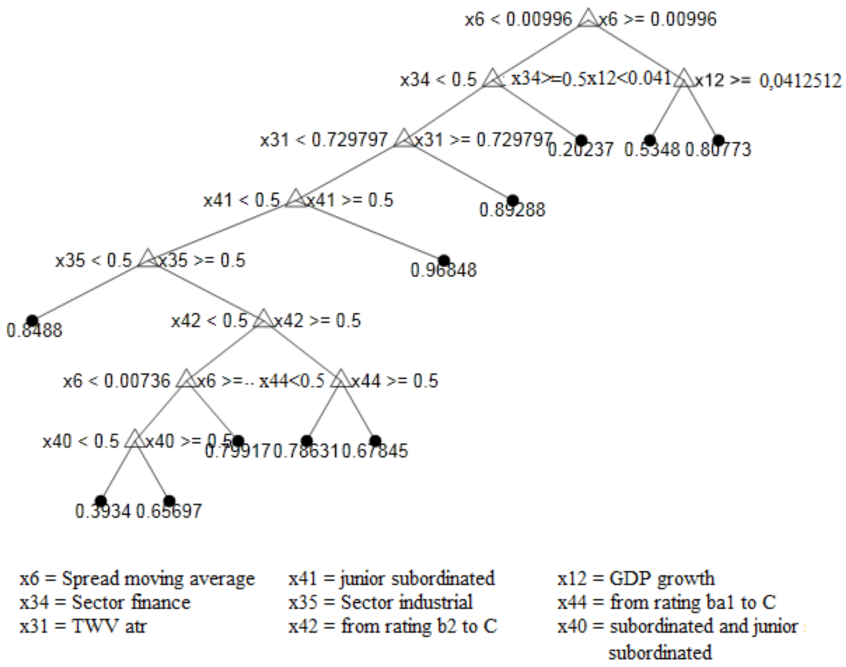


Figure 6: Regression tree HELOC out-of-time

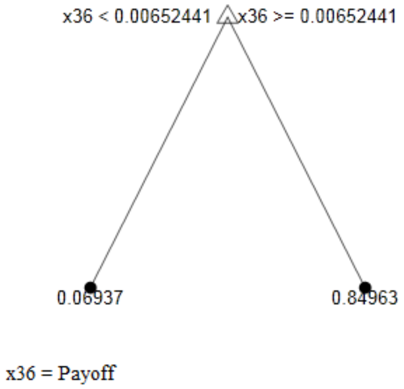
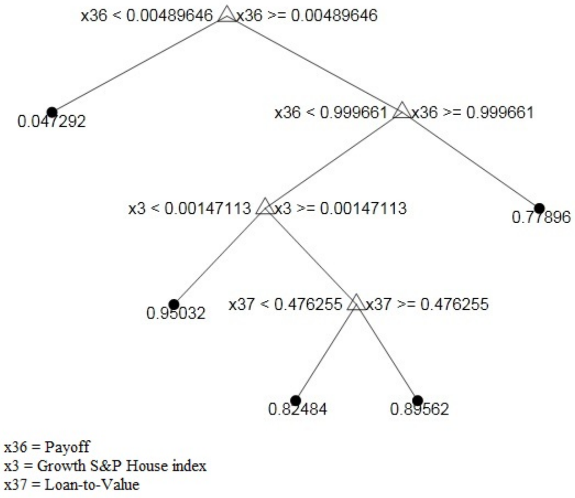


Figure 7: Regression tree HELOC out-of-sample



more accurate LGD prediction, the results suggest that financial institutions should use a two-stage model. The incorporation of macroeconomic variables improves the predictive performance of the models, thus confirming the relationship between the business cycle and LGD found in previous studies. The variables incorporated differ between the models and datasets. It is thus recommended that financial institutions estimate a different model for each of their portfolios. It is important that banks select the right variables to estimate downturn LGD. We found that the corporate LGD is more responsive to higher default periods than home equity loan losses. However, the long-run default weighted average of corporate losses is significantly lower than the average of losses on home equity loans.

We found that the response of LGD to a downturn period differs amongst the techniques considered. The two stage models appears to be the most responsive for two out of the three[4] forecasts conducted in this research.

Acknowledgements

Bart Baesens acknowledges the Flemish Research Council for financial support (Odysseus grant B.0915.09). We would like to thank the financial institutions who provided us with the LGD data.

References

[1] Altman E, Brady B, Resti A, Sironi A (2005).*The PD/LGD link: empirical evidence from the bond market.* In: Altman E, Resti A, Sironi A (eds) Recovery risk. Riskbooks, London, pp 217233

[2] Basel Committee on Banking Supervision, (2004). *Background note on LGD quantification*, Bank for International Settlements, 6 December 2004.

[3] Basel Committee on Banking Supervision (2006), *International Convergence of Capital Measurement and Capital Standards.* Basel, Bank for International Settlements.

[4] Bastos, JA (2009). *Forecasting bank loans for loss-given-default.* CEMAPRE working papers 0901. Centre for Applied Mathematics and Economics. School of Economics and Management. ISEG. Technical University of Lisbon.

[5] Bellotti, T. and Crook, J. (2011). Loss given default models incorporating macroeconomic variables for credit cards. *International Journal of Forecasting*, 28,171-182.

[6] Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees.* Chapman & Hall/CRC.

[7] Box, G., & Cox, D. (1964). An analysis of transformations. *Journal of the Royal Statistical Society*, Series B, 26, 211-243.

[8] Caselli,S. , Gatti, s. & Querci, F., (2008). 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.

[9] Cohen, P., Cohen, J., West, S. G., & Aiken, L. S. (2002). *Applied multiple regression/correlation analysis for the behavioral sciences.* Lawrence Erlbaum.

[10] Dermine, D., & de Carvalho, C. N. (2005). Bank loan losses-given-default: a case-study. *Journal of Banking and Finance*, 30(4), 12191243.

[11] Draper, N., & Smith, H.(1998). *Applied regression analysis.* Wiley.

[12] Loterman, G., Brown, I., Martens, D., Mues, C., & Baesens, B. (2011). Benchmarking regression algorithms for loss given default modeling. *International Journal of Forecasting*, 28,161-170.

[13] Martens, D., Baesens, B., Van Gestel, T., & Vanthienen,J. (2007). Comprehensible credit scoring models using rule extraction from Support Vector Machines. *European Journal of Operational Research*,183,1466-1476.

[14] Schuermann, T. (2004). *What do we know about Loss Given Default?.* Wharton Financial Institutions Center Working Paper No. 04-01.

[15] Van Gestel, T., Suykens, J., Baesens, B., Viaene, S., Vanthienen, J., Dedene, G., et al. (2004). Benchmarking least squares Support Vector Machine classifiers. *Machine Learning*, 54, 5-32.

[16] Van Gestel, T., Baesens, B., Van Dijcke, P., Garcia, J., Suykens, J. & Vanthienen, J. (2005). A process model to develop an internal rating system: sovereign credit rating. *Decision Support Systems*,Vol. 42, Issue 2 ,1131-1151.

[17] Van Gestel, T., Martens, D., Baesens, B., Feremans, D., Huysmans, J. & Vanthienen,J., (2007). Forecasting and analyzing insurance companies' ratings. *International Journal of Forecasting*, 23,513-529.

[4] This refers to the out-of-time forecast of the HELOC dataset and the out-of-sample forecast conducted on both datasets