Machine Learning in Credit Risk Management: An Empirical Analysis for Recovery Rates

Zur Erlangung des akademischen Grades eines

Doktors der Wirtschaftswissenschaften

(Dr. rer. pol.)

von der Fakultät für Wirtschaftswissenschaften

des Karlsruher Instituts für Technologie

genehmigte

DISSERTATION

von

M. Sc. Konstantin Heidenreich

Tag der mündlichen Prüfung: 18.12.2018

Referent: Prof. Dr. Gholamreza Nakhaeizadeh

Korreferentin: Prof. Dr. Melanie Schienle

Schweinfurt, im Oktober 2018

Acknowledgements

First of all, I would like to express my sincere gratitude to my supervisors Prof. Dr. Gholamreza Nakhaeizadeh and Prof. Dr. Melanie Schienle for their valuable comments and for taking on the role as my supervisors. I would like to thank Prof. Frank J. Fabozzi for his extremely helpful comments for the three papers which chapters 2, 3, and 4 are based on. I would also like to thank Prof. Uhrig-Homburg for providing the recovery dataset used in Chapter 2 and Chapter 3. I would especially like to thank Dr. Abdolreza Nazemi. His insightful comments, encouraging support and his guidance during the process of writing this dissertation helped me a lot. I dedicate this dissertation to my grandmother Elfriede Heidenreich. Finally, I would like to thank my girlfriend Julia and my mother as well as my family and friends for their love and support.

Contents

List of Figures

List of Tables

List of Abbreviations

AIC	Akaike information criterion
BIC	Bayesian information criterion
BBG	Bloomberg
CART	Classification and regression technique
CPI	Consumer price index
CVaR	Conditional Value-at-Risk
DB	With different biases
DE	Differential evolution
FRED	Federal Reserve economic data
FX	Foreign exchange
GDP	Gross domestic product
IRB	Internal rating based
ISM	Institute for supply management
KPCA	Kernel principal component analysis
LASSO	Least absolute shrinkage and selection operator
LGD	Loss-given-default
Lin. reg.	Linear regression
LS-SVM	Least-squares support vector machine
LS-SVR	Least-squares support vector regression
MAE	Mean absolute error
MSE	Mean squared error

NLPCA	Nonlinear principal component analysis
OLS	Ordinary least squares
PEP	Power expectation propagation
PC	Principal component
PCA	Principal component analysis
PD	Probability of default
PPI	Producer price index
R^2	Coefficient of determination
Reg. tree	Regression tree
RF	Random forest
RMSE	Root mean squared error
RR	Recovery rate
RVI	Relative variable importance
SenSec	Senior secured
SenSub	Senior subordinated
SenUn	Senior unsecured
SP	Semiparametric
S&P	Standard & Poor's
SPCA	Sparse principal component analysis
STD	Standard deviation
Subord	Subordinated
SVM	Support vector machine
SVR	Support vector regression
TRACE	Trade reporting and compliance engine
VaR	Value-at-Risk

Chapter 1

Introduction

In previous credit risk analyses, for example in a study by Virolainen (2004), recovery rates have generally been assumed to be constant. However, the variation in the recovery rates during the Great Financial Crisis of 2007 and 2008 has shown that this assumption of a constant recovery rate is unrealistic. In the aftermath of the 2007/2008 financial crisis financial regulation has been enhanced by the implementation of the Basel accords.

The Basel II accord (Basel Committee on Banking Supervision, 2006) allows for three different approaches in terms of calculating the credit risk exposures: (1) a standardized approach using the credit ratings assigned by credit rating agencies to compute a risk-weighted asset basis for determination of the required minimum capital, (2) a foundation internal ratings-based approach that allows that allows for the use of internal estimates of the probability of default or (3) an advanced internal ratings-based approach that uses internal estimates for the three risk parameters that relate to credit risk exposure. Financial institutions using the advanced internal ratings-based approach need to develop proprietary methods for estimating the key risk parameters which include loss-given-default, the exposure at default and the probability of default.

Among the key elements of the Basel regulations are more advanced capital requirements for financial institutions. As outlined by Nazemi and Heidenreich (2017), the required equity capital resulting from the determination of the capital requirements is expensive for financial institutions to hold. Thus, more precise estimates of the credit risk parameters can yield a significant economic value for financial institutions. Further, more accurate recovery rate predictions are needed not only for calculating

regulatory capital requirements but also for estimating economic capital requirements and would furthermore enable a more accurate pricing of financial instruments for trading and investment purposes.

While much attention has been paid to the probability of default in the literature, the recovery rates have thus far been less well examined. Thus, the focus of this dissertation is placed on the estimation of the recovery rates of U.S. corporate bonds, estimates that would allow financial institutions and regulators to gain insights into and to understand the modeling of recovery rates.

Altman and Kishore (1996) have stated that the industries with the highest average recovery rates are the public utilities and the chemical sectors. Furthermore, they report that the size of the default issue, the time between issue and the default date, and the original bond rating do not have a significant effect on recovery rates. Altman et al. (2005) demonstrate a close relationship between the default rates and the recovery rates, with Acharya et al. (2007) presenting evidence that defaulted firms exhibit significantly lower recovery rates if the related industry is in distress. Further, they show that defaulted firms in distressed industries are more likely to be restructured than to be acquired or liquidated. In a study by Jankowitsch et al. (2014) bonds in a formal legal bankruptcy procedure are shown to exhibit lower recovery rates than the bonds of firms undergoing an out-of-court restructuring. In addition, the same authors demonstrated that senior bonds have, on average, higher recovery rates than subordinated debt. Moreover, they find that the average recovery rate exhibits substantial variation over time. Altman and Kalotay (2014) condition mixtures of Gaussian distributions on instrument characteristics, borrower characteristics and credit conditions, an approach that generates forecasts which are more accurate than parametric regression-based methods during out-of-time estimation.

Qi and Zhao (2011) report that non-parametric methods, such as regression trees and neural networks, outperform parametric methods. In a comparative study by Loterman et al. (2012), non-linear techniques, such as neural networks and support vector machines, were shown to perform significantly better than traditional linear techniques. Yao et al. (2015) apply support vector methods to the prediction of the recovery rates of corporate bonds. They show that support vector methods outperform more traditional modeling techniques such as linear regression. Moreover, they

further report that predictive accuracy is not increased by standard transformations of the recovery rate such as beta-transformation or log-transformation. Tobback et al. (2014) show that incorporating macroeconomic variables improves predictive performance significantly using two data sets relating to home equity and corporate loans. Kalotay and Altman (2017) emphasize the importance of accounting for time variation in recovery rate prediction and that, in particular, they show that conditional Gaussian mixture models yield improved recovery estimates. Further, they outlined how the best instrument-level forecasts often miss the association between default probability and recovery rates, making them less suitable at the portfolio level.

Inspired by the good performance of non-parametric methods, such as regression trees and support vector methods, discussed in the literature, further research relating to the performance of these techniques would appear to be timely. In particular, least-squares support vector regression, two further variants of least-squares support vector regression, ϵ-insensitive support vector regression and regression trees will be examined. By comparing the various performance measures determined for each model, further insights into a particular models suitability for recovery rate prediction can be obtained. As the performance benchmark, a traditional linear regression model is used.

A comprehensive review of the recovery rate literature may also reveal insights into related areas, such as the recovery rate modeling relating to other assets such as bank loans or credit card debt and modeling of the probability of the default of corporate bonds. Based on a review of the literature, relevant explanatory variables will be identified and data relating to defaulted corporate bonds will be collected, cleaned up and matched to the corresponding variables. An exploratory analysis of the data set will also be presented. Additionally, the resulting coefficients of the linear regression model will be studied with respect to their economic interpretation and the methodology for choosing appropriate hyperparameters, used in relating to the machine learning techniques, will be outlined.

In a second step, the issue of whether incorporating information from a broad range of macroeconomic variables has increased predictive power compared to the literature, in which only a few macroeconomic variables have been taken into account, will be investigated. Data reduction techniques, such as principal component analysis (PCA), autoassociative neural networks, kernel PCA and sparse PCA will be studied in order to

extract the macroeconomic factors. Moreover, methods such as gradient boosting (for ranking the macroeconomic variables) and the least absolute shrinkage and selection operator (for selecting the macroeconomic variables) will be investigated. Techniques such as stability selection will be applied to account for multicollinearity during the selection of the macroeconomic variables.

Another research question relates to whether the predictive performance of the machine learning techniques used in the literature can be increased by fusing the outputs from multiple models based on fuzzy rules. We create a fuzzy rule base with a differential evolution algorithm which is able to cope with complex data and which allows us to avoid the difficulties associated with higher dimensionality. Regarding the defuzzification, we compare the performance of the maximum formula, the maximum of maxima formula, the mean formula and the mean of maxima formula, and examine whether the performance of the fuzzy models improves through the addition of the principal components of 104 macroeconomic variables. In addition, the performance of our prediction techniques after normalizing the macroeconomic variables with the Box-Cox transformation will be examined.

The usefulness of machine learning techniques for out-of-time prediction will be investigated. With the exception of Altman and Kalotay (2014) and Kalotay and Altman (2017), most studies have compared the performance of different modeling techniques during cross-validation or through testing on a random portion of the dataset out-of-sample. For practical purposes however, predictive capacity out-of-time is essential, so that the out-of-time performance of the machine learning techniques will be compared to more traditional statistical techniques. Furthermore, to the best of our knowledge, there is no study which examines the importance of the explanatory variables in a high-dimensional analysis for recovery rate prediction. Therefore, the determination of which groups of variables add a greater predictive power to our models will allow us to generate insights into what the actual drivers of recovery rates are. Finally, due to the regulatory requirements illustrated above, it is interesting to examine the behaviour of the recovery rates during times of macroeconomic stress.

This thesis contributes to the existing literature on recovery rate prediction of corporate bonds in several ways. In Chapter 2, we contribute to the credit risk literature in four ways. Whilst in the literature, such as in the work of Jankowitsch et al. (2014),

only a small set of macroeconomic variables is utilized, we include of a broad range of macroeconomic variables in our multi-factor framework. We show that including the principal components of the macroeconomic variables relating to a wide range of categories, such as lending conditions, micro-level conditions, business cycle conditions, stock market conditions and international competitiveness, improves the predictive performance of our models. Using sparse principal component analysis instead of principal component analysis, the predictive capacity of our models is further increased and the principal components are more easily interpretable. On examining the relative importance of macroeconomic variables for recovery rate prediction with gradient boosting, we find that the credit spread of corporate bonds, the yields offered on corporate bonds, the annual return of the Russell 2000 and the number of unemployed, are the most informative variables. Adding the 20 most important macroeconomic variables from our ranking by relative importance (with gradient boosting), we improve the performance of easy to interpret models such as the linear regression and regression tree models.

Three contributions to the literature are made in Chapter 3. We present the first study that applies fuzzy decision fusion models to corporate bond recovery rate prediction and show that adding the principal components from 104 macroeconomic variables improves the predictive accuracy of our models. Further, we test the application of the Box-Cox transformation as a potential means of increasing the predictive capacity of our models and find that our fuzzy models outperform previously suggested techniques, such as the three variants of least-squares support vector regression, regression trees and linear regression approaches.

In Chapter 4 we make four contributions to the existing credit risk literature. We use high-dimensional data and we compare techniques, such as the stability selection, the SparseStep algorithm and the MC+ algorithm, for selecting the most important macroeconomic variables. We take alternative independent variables, such as text-based variables, into account for our analysis. In contrast to the literature, such as discussed in the work of Kalotay and Altman (2017), we find that machine learning techniques, such as least-squares support vector regression, random forest, regression tree and a sparse Gaussian process approximation using power expectation propagation, outperform more traditional techniques such as linear regression and inverse Gaussian regression out-of-sample and in intertemporal analysis. We examine the permutation importance of the following groups of independent variables in predicting

recovery rates: seniority variables, industry variables, bond characteristics, news-based measures, financial conditions, monetary measures, corporate measures, business cycle measures, stock market conditions, international competitiveness and micro-level conditions.

The remainder of this dissertation is based on three independent research papers. In Chapter 2, the first paper (Improving corporate bond recovery rate prediction using multi-factor support vector regressions) is presented. Chapter 3 concerns the second paper (Fuzzy decision fusion approach for loss-given-default modeling). The third paper (Intertemporal defaulted bond recovery prediction via machine learning) is the subject of Chapter 4. We conclude with a summary of our results and a discussion relating to potential further research.

Chapter 2

Improving corporate bond recovery rate prediction using multi-factor support vector regressions

This chapter is joint work with Dr. Abdolreza Nazemi[1] and Prof. Frank J. Fabozzi[2] published in 2018 as: Improving corporate bond recovery rate prediction using multi-factor support vector regressions, European Journal of Operational Research, 271(2), 664-675. https://doi.org/10.1016/j.ejor.2018.05.024

2.1 Introduction

Extended regulation of the financial industry as set forth in the Basel accords has focused on the imposition of stricter (i.e., higher) capital requirements. According to Schuermann (2004), calculation of expected loss is the product of three measures: exposure at default, the probability of default, and the loss given default. Though the probability of default has been the main focus of practitioners and researchers for calculating the minimum capital requirement, loss given default has been comparatively less investigated. As a result of the Basel II accord, the importance of loss given default, however, has increased substantially for banks and other financial institutions.

According to Loterman et al. (2012) the impact of loss given default on the required minimum capital is linear within the required framework of Basel II. Improved

[1] School of Economics and Business Engineering, Karlsruhe Institute of Technology
[2] EDHEC Business School, Nice, France

prediction models facilitate the calculation of more reasonable capital requirements and enable more precise valuations of defaulted bonds for trading purposes. Equivalently, since one minus the loss given default is the recovery rate, the focus in this paper is on the recovery rate.

Traditionally, linear regression has been applied to predict recovery rates. Altman and Kishore (1996) document that average recovery rates from utility companies and chemical companies are significantly higher than in other industries. Cantor and Varma (2004) study the determinants of recovery rates and find out that seniority and security are the two most important exploratory variables. Exploring the relationship between recovery rates and aggregate default rates, Altman et al. (2005) conclude that recovery rates of corporate bonds are related to default rates, seniority and collateral levels. Acharya et al. (2007) investigate how the distress of the industry of a defaulted firm affects the recovery rate. A beta regression model to predict recovery rates of bank loans is suggested by Calabrese and Zenga (2010). Bastos (2010) reports that the predictive accuracy for regression trees is higher than for parametric models. Rösch and Scheule (2014) propose a joint estimation approach for probabilities of default and recovery rates. Altman and Kalotay (2014) suggest an approach to model the distribution of recovery rates based on mixtures of Gaussian distributions conditioned on borrower characteristics, instrument characteristics and credit conditions. Their method outperforms parametric regressions as well as regression trees.

According to Jankowitsch et al. (2014) the significance of calculating precise forecasts of loss given default has been increased by the variability and volume of defaults during the financial crisis of 2007/2008. Imprecise forecasts might have been undetected and might have born less risk as less default events occurred.

In the paper, we use various factors such as instrument-specific characteristics, industry-specific variables, and macroeconomic variables to forecast recovery rates. Although traditional regression analysis has been used in the literature to project recovery rates, two studies suggest that alternative statistical models can improve forecasts. Loterman et al. (2012) compare 24 techniques for the prediction of recovery rates of various debt instruments such as mortgage loans, corporate loans, and personal loans. They show a clear tendency that non-linear models such as artificial neural networks and support vector machines have a higher predictive capacity than traditional linear

techniques. Furthermore, they conclude that the predictive power of two-stage models – a combination of non-linear and linear models – is on par with the predictive capacity of non-linear models while two-stage models are more easily interpretable due to the linear part of the model. Yao et al. (2015) forecast recovery rates of corporate bonds using support vector techniques. They report that applying three alterations of a least-squares support vector regression (LS-SVR) significant outperformance versus traditional modeling techniques such as fractional response regression or linear regression is observed. Further, they argue that LS-SVR outperforms compared to traditional approaches when segmenting the bonds by their seniority.

Motivated by the findings of Loterman et al. (2012) and Yao et al. (2015), we use support vector regression (SVR) models to determine if the forecasts of these models improve the forecasts relative to traditional linear regression analysis. In contrast to linear regression, SVR allows one to model non-linearities by employing a non-linear kernel function. The independent variables are implicitly mapped from the low-dimensional input space into a high-dimensional feature space via the kernel function. By doing so, the kernel function needs not be calculated explicitly. After mapping to the high-dimensional linear space, SVR can provide more accurate predicts. The four SVR models used are an ϵ-insensitive SVR, a LS-SVR, and two modified LS-SVR methods to account for heterogeneity within the seniority classes. The out-of-sample forecasts from these four SVR models are then compared to assess whether these models can outperform the forecasts obtained from traditional regression analysis.

In addition to the use of alternative models to the traditional regression analysis, we use a more extensive set of macroeconomic variables to forecast recovery rates. Our suggested recovery rate models for U.S. corporate bonds contribute to Yao et al. (2015) in several ways. They utilize only a small set of macroeconomic variables whereas we make use of a broad range of macroeconomic variables with applying multi-factor SVR. We compare the predictive performance using different data reduction techniques for the 104 macroeconomic variables such as principal component analysis (PCA), SPCA, NLPCA, KPCA and gradient boosting. Further, we investigate the relative importance of these macroeconomic variables with gradient boosting to generate a ranking of the macroeconomic variables.

Tobback et al. (2014) highlight the relevance of macroeconomic independent vari-

ables when forecasting recovery rates of corporate loans and home equity loans. They apply a linear regression, a regression tree, SVR and a two-stage model merging the linear model with SVR. Including 11 macroeconomic variables improves the performance of these models significantly. Moreover, Duffie et al. (2009) and Koopman et al. (2011) argue that macroeconomic influences matter substantially for forecasting the probability of default. Specifically, they show the influence of a latent and dynamic frailty factor. Koopman et al. (2011) include the first 10 principal components of more than 100 macroeconomic variables in their analysis to ensure that the fraily factors represents only truly unobservable effects.

Numerous studies have considered a limited number of financial and macroeconomic variables for the prediction of recovery rates. Most recovery rate research has applied statistical or machine learning models, which cannot handle a large number of predictors. For example, we need to iterate a stepwise-regression for 2^{104-1} times for selecting the best set of macroeconomic variables, which is empirically impossible. Because data-reduction techniques overcome this limitation we introduce four types of principal component analysis techniques and the gradient boosting model to the recovery rate modeling research. We merge 104 macroeconomic variables with bond-specific data. We apply PCA to 104 macroeconomic variables capturing 96% of the dataset's variance in our analysis as variables in our models. Alternatively, we apply sparse PCA, nonlinear PCA from an autoassociate neural network, and kernel PCA to obtain their principal components. To the best of our knowledge, this is the first study comparing different PCA techniques in credit risk analysis. In addition, we apply gradient boosting to determine the relative importance of the macroeconomic variables in our analysis and to enable a ranking of the 104 macroeconomic variables for recovery rate prediction.

We study the performance of machine learning techniques such as ϵ-insensitive SVR, regression tree and three variants of LS-SVR in comparison to a more traditional linear regression approach. In particular, we include information from an extensive set of macroeconomic variables in our analysis. Beyond that, we compare data reduction techniques such as PCA, SPCA, NLPCA, KPCA, and gradient boosting to achieve dimensionality reduction of the 104 macroeconomic variables. We have organized the paper as followed. An outline of our multi-factor framework and the variables selected is provided in the next section, Section 2.2.2. We also present the data-reduction techniques we apply to the 104 macroeconomic variables.

We give a description of our choices of modeling techniques which are linear regression, regression tree and SVR in Section 2.3. In Section 2.4, we present an exploratory analysis of our dataset which consists of 775 corporate bonds with default events between 2002 and 2012. In the middle part of Section 2.4 we demonstrate the increased predictive accuracy of LS-SVR while comparing the out-of-sample performance of the cross-validated models. Specifically, the addition of the principal components of 104 macroeconomic variables increases the models' performance. We rank the macroeconomic variables applying gradient boosting and discuss the predictive performance when adding the highest ranked macroeconomic variables in the final part of Section 2.4. We conclude our paper in Section 2.5.

2.2 Multi-factor framework

In the following section we give a description of our multi-factor framework, its independent variables and its expansion by adding the principal components of macroeconomic variables.

2.2.1 Selecting factors for modeling

In our framework the recovery rate r_{ij} for bond i of firm j is defined as follows:

$$r_{ij} = \alpha + \beta_c X_{ci} + \beta_m X_{mi} + \beta_{ind} X_{indj} + \epsilon_{ij} \tag{2.1}$$

where

X_{ci} denotes a vector of instrument characteristics of bond i;

X_{indj} is a vector with the industry characteristics of the defaulted firm j, and;

X_{mi} is a vector of the macroeconomic variables for the year preceding the default of the i-th observation.

The instrument characteristic included in the model are the bond's seniority in the capital structure, the amount of the bond's trading volume, and a dummy variable indicating defaults under Chapter 11 versus bonds that have been assigned a rating

that is the equivalent of a default. For industry characteristics, we include a dummy variable for the utility industry and two variables measuring whether the industry is in distress. The first industry distress dummy variable is based on whether the performance of the industry index was worse than -30% in the year preceding the default. The second industry distress dummy variable is based on whether the sales growth in the respective industry in the year preceding the default was negative. The macroeconomic variables include (1) the number of defaulted bonds in the respective year, (2) the value of a high-yield index, (3) the change in gross domestic product (GDP), (4) the unemployment rate, and (5) the federal funds rate. The macroeconomic variables are observed in the year preceding the default

The selection of variables is based on a thorough overview of the literature on the determinants of recovery rates. The study by Altman and Kishore (1996) is one of several studies that argue that recovery rates vary across industries. So, we include industry dummies for the utility industry, the financial industry, the communications industry, the cyclical consumer goods industry and the industrial industry. Among others Cantor and Varma (2004) report an increased predictive capacity by taking the seniority class of the debt instrument into account. If a default event occurs the claims of the most senior bondholders are the first to be paid off while the claims of the junior/subordinated bondholders are the last be paid off. (Although studies find that the absolute priority principle does not hold in the case of a Chapter 11 bankruptcy, senior creditors do generally fair better than subordinated creditors.) According to Acharya et al. (2007) both the performance of an industry index and the sales growth of the industry explain part of the variation in recovery rates. Moreover, Acharya et al. (2007) have shown how industry distress dummy variables based on these two independent variables have a statistically significant effect on the recovery rate. For example, an industry that is in structural decline because of its reliance on an outdated technology might experience negative impact on the recovery rates of the firm in that industry.

We use a variable found to be significant by Altman et al. (2005) as a proxy for the volume of defaulted bonds: the number of defaulted bonds in the default year. The current level of a high yield index is used as a proxy for the aggregated market default rate, a significant independent variable reported by Cantor and Varma (2004). The liquidity of the bond is a significant variable according to Jankowitsch et al. (2014); so as a proxy for liquidity we include the aggregated trade volume of the defaulted bonds

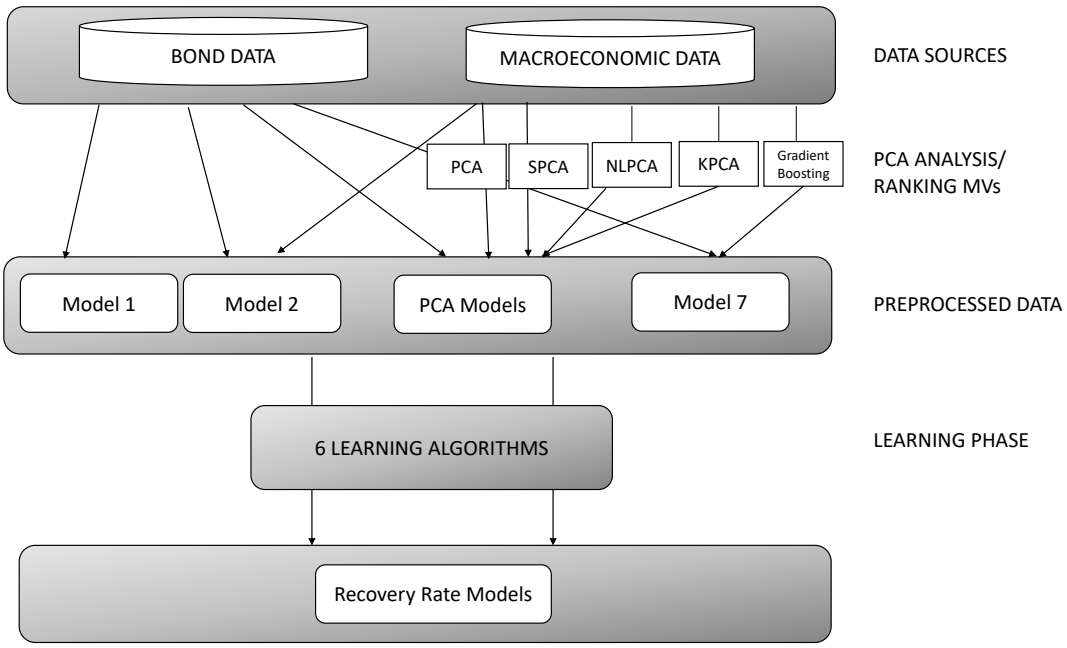

Figure 2.1: The steps for predicting the recovery rates

in the 30 days after the default. This corresponds to a liquidity premium.

2.2.2 Extension of the framework

As shown in Figure 2.1, we extended the basic framework of Model (1) by including the principal components of 104 macroeconomic variables from a broad range of categories such as stock market conditions, credit market conditions, international competitiveness, business cycle conditions, and micro-level conditions. The 104 macroeconomic variables shown in Table 2.1 have been observed in the year preceding the respective default. These principal components substitute the macroeconomic variables included in Model (2). In using that many macroeconomic variables we differ from other studies using only a small number of macroeconomic variables for LGD modeling such as Tobback et al. (2014).

2.2.2.1 Principal component analysis

For implementation purposes we use the dimensionality reduction toolbox from Van der Maaten et al. (2007). The macroeconomic variables that are used to generate the principal components are denoted as X_{pcai} with dimension 104×1. They can be

Table 2.1: Macroeconomic variables for principal components analysis

Financial Conditions	
Total real estate loans	Household obligations/income
Federal debt of non-fin. industry	Total loans and leases, all banks
Total commercial loans	Non-performing loans ratio
Total consumer credit outst.	Non-performing loans ratio small banks
Commercial & industrial loans	Net loan losses
Excess reserves of dep. institutions	Total net loan charge-offs
Total borrowings from fed reserve	Return on bank equity
Bank loans and investments	Loan loss reserves
Household debt service payments	Non-perf. commercial loans
Business cycle indicators	
GDP growth	Civilian employment
ISM manufacturing index	Employment/population-ratio
Industr. production index	Unemployed, more than 15 weeks
Uni Michigan consumer sentiment	New orders: durable goods
Private fixed investments	Final sales of dom. product
Real disposable personal income	New orders: capital goods
National income	Inventory/sales-ratio
Personal Income	Capacity util. manufacturing
Manuf. industry output	Change in private inventories
Consumption expenditure	Capacity util. total industry
Manuf. industry production	Inventories: total business
Expenditure durable goods	Light weight vehicle sales
Government expenditure	Non-farm housing starts
Gross private domestic investment	Housing starts
Unemployment rate	New houses sold
Total no unemployed	New building permits
Weekly hours worked	Final sales to domestic buyers
Stock Market Indicators	
S&P 500	Russell 2000
S&P 500 Vol	Russell 2000 Vol
Dow Jones industrial average Vol	S&P small cap index
Nasdaq 100	S&P small cap index Vol
Nasdaq 100 Vol	
International Competitiveness	
Trade weighted USD (Dollar Index)	Real exports goods, services
FX index major trading partners	Balance on merchandise trade
Current account balance	Real imports goods & services
Micro-level factors	
Unit labor cost: manufacturing	Effective federal funds rate
Unit labor cost: nonfarm business	Corporate yield spread (baa seasoned bonds)
Total wages & salaries	AAA corporate bond yield
Non-durable manufacturing wages	30 year mortgage rate
Management salaries	BAA corporate bond yield
Durable manufacturing wages	Volume defaulted bonds
Employment cost index: benefits	PPI all commodities
Employee compensation index	PPI industrial commodities
Employment cost index: wages & salaries	PPI interm. energy goods
1 month commercial paper rate	PPI crude energy materials
Treasury bond yield, 10 years	PPI finished goods
3 month commercial paper rate	PPI intermediate materials
Term structure spread	
Monetary measures	
M2 money stock	Personal savings
CPI: all items less food	Personal savings rate
Uni Michigan infl. expectations	Gross saving
CPI: energy index	GDP deflator, implicit
Corporate measures	
Corp. profits	Net corporate dividends
After tax earnings	Corporate net cash flow

presented in the following structure:

$$X_{pcai} = \Lambda FPCA_i + \zeta_i, \qquad i = 1, ..., N. \tag{2.2}$$

We estimate the vector of factors $FPCA_i$ by using principal components analysis (PCA) where Λ are the loadings.

Following Koopman et al. (2011) the factors $FPCA_i$ from the observed macroeconomic variables are estimated by minimizing the objective function:

$$\min_{FPCA,\Lambda} V(FPCA, \Lambda) = (N)^{-1} \sum_{i=1}^{N} (X_{pcai} - \Lambda FPCA_i)'(X_{pcai} - \Lambda FPCA_i) \tag{2.3}$$

We have standardized and normalized the macroeconomic variables to obtain an unconditional unit variance for $n = 1, ..., N$. The observed macroeconomic variables are denoted by a 104×1 vector $X_{pcai} = (x_{pcai1}, ..., x_{pcai104})'$. Using $S_{X'X} = N^{-1} \sum_i X_{pcai} X'_{pcai}$ as the sample covariance, equation (2.3) can be transformed to a maximization problem of the following form:

$$\begin{aligned} &\max_{\Lambda} \; tr(\Lambda' S_{X'X} \Lambda) \\ &\text{s.t. } \Lambda'\Lambda = I_R \end{aligned} \tag{2.4}$$

With $\hat{\Lambda}$ as normalized eigenvectors for the R largest eigenvalues of $S_{X'X}$ the principal components estimator results as:

$$\hat{FPCA}_i = X'_{pcai} \hat{\Lambda} \tag{2.5}$$

The independent variables $\hat{FPCA}_i$ that are found to capture more than 96% of the variance of the observed macroeconomic variables, are then used as inputs to our models in the empirical analysis. We replaced the five single macroeconomic variables with the principal components of 104 macroeconomic variables. So, by including the principal components of the 104 macroeconomic factors our framework is extended to:

$$r_{ij} = \alpha + \beta_c X_{ci} + \beta_{ind} X_{indj} + \beta_{pi} \hat{FPCA}_i + \epsilon_{ij} \tag{2.6}$$

2.2.2.2 Sparse principal component analysis

Having 104 macroeconomic variables with non-zero factor loading in PCA makes it difficult to interpret the economic meaning of each PC. Fortunately, elastic net regularization of PCA, as outlined by Zou et al. (2006), produces a more robust estimate of factors loadings. By applying an L_1 and an L_2 penalty to the coefficients of the PCA regression formulation, sparse principal components can be estimated. Therefore, we use SPCA as presented by Zou et al. (2006) to identify the influence of our 104 macroeconomic variables. We implement SPCA using the SpaSM toolbox by Sjöstrand et al. (2012). Effectively, we have a regression-like framework of PCA with an elastic net regularization:

$$\begin{aligned} \{\hat{\Lambda}, \hat{B}\} = \arg\min_{\Lambda,B} \|X - XB\Lambda'\|^2 + \delta\|B\|^2 + \lambda\|B\|_1, \\ \text{s.t. } \Lambda'\Lambda = I_R \end{aligned} \tag{2.7}$$

where λ denotes the Lasso regularization coefficient and δ denotes the Ridge regularization coefficient. B denotes the elastic net regularization loadings and X_{pcai} denotes the 104×1 vector $X_{pcai} = (x_{pcai1}, ..., x_{pcai104})'$ of macroeconomic variables.

Denoting $\hat{\Lambda}$ as normalized eigenvectors for the R largest eigenvalues from the PCA solution in 2.1.2, we calculate the estimated value of the independent variables, denoted by $\hat{SPCA}$, as follows:

$$\hat{SPCA}_i = X'_{pcai}\hat{B}'\hat{\Lambda} \tag{2.8}$$

2.2.2.3 Nonlinear principal component analysis

To account for nonlinear relationships between the 104 macroeconomic variables we also investigate the performance of nonlinear principal component analysis (NLPCA). One class of NLPCA that has proved successful is autoassociative neural networks as introduced by Kramer et al. (1991). Building on this work, Hsieh et al. (2007) introduce an inconsistency index as information criterion to avoid overfitting in the choice of hyperparameters.

An autoassociative neural network has one input layer, three hidden layers, and one output layer. Hence, the number of hidden neurons in the middle hidden layer is restricted to the number of bottleneck neurons, which are effectively the principal components. The mean squared error (MSE) between the input layer and the output layer is minimized. In the process, the nonlinear principal components in the middle

hidden layer, that is the bottleneck layer, are calculated.

We implement the autoassociative neural network for NLPCA with the neumatsa toolbox by Hsieh et al. (2006). To avoid overfitting we use the inconsistency index as described by Hsieh et al. (2007) as the information criterion for cross validation to determine the weight decay and the number of hidden neurons in the first hidden layer.

2.2.2.4 Kernel principal component analysis

Another nonlinear dimensionality reduction technique is kernel principal component analysis (KPCA) as introduced by Schölkopf et al. (1997). An advantage of KPCA in comparison to autoassociative neural networks and other methods for nonlinear PCA is that no nonlinear optimization is necessary for KPCA. So, there is no danger of finding a solution that is a local minimum because KPCA constitutes an eigenvalue problem as in a standard PCA.

According to Schölkopf et al. (1997), KPCA uses a nonlinear transformation $\phi(X)$ from the original feature space to a higher-dimensional feature space. Then, standard PCA is performed in the higher-dimensional feature space. The calculation does not require one to explicitly calculate the mapping $\phi(X)$ but only requires one to calculate the dot product $\phi(X)\phi(X)^T$. Consequently, any kernel function such as polynomial kernels, radial basis functions, and sigmoid kernels can be used for the implicit mapping. We process our data using the KPCA toolbox from Wang (2012). We choose the radial basis function as kernel.

2.3 Corporate bond recovery rate modeling

In this section we present the models we use for the empirical analysis reported in Section 2.4. In what follows, we let X denote the matrix of features that is composed of the vector of instrument characteristics X_{ci}, the vector of industry characteristics of the defaulted firm X_{indj}, and either the vector of macroeconomic variables X_{mi} (Model (2)) or the vector of principal components of a large set of macroeconomic variables (Models (3)-(6)).

2.3.1 Support vector regression

As shown by Bellotti and Crook (2009) and Danenas and Garsva (2009), support vector machines are a helpful tool in the domain of credit risk. To the best of our knowledge only one study, Yao et al. (2015), has attempted to predict corporate bond recovery rates using support vector techniques. As outlined by Chalup and Mitschele (2008), support vector techniques are promising techniques for finance applications because of their ability to deal with non-linear input data. So, LS-SVR as a "kernelized" version of the traditional linear regression might yield a higher predictive capacity.

As shown by Aizerman et al. (1964), Mercer's theorem allows a computationally efficient calculation of a kernelized problem. To make use of Mercer's Theorem an appropriate kernel function has to be chosen. As stated by Chalup and Mitschele (2008), the only prerequisites a kernel has to fulfill are to be positive semi-definite and to represent a similarity measure between pairs of input samples. More specifically, we use the Radial Basis Function kernel in the following form:

$$K(X_i, X_j) = \exp\left(-\frac{||X_i - X_j||^2}{2\sigma^2}\right) \tag{2.9}$$

2.3.1.1 Least-squares support vector regression

We use three different LS-SVR models. The first one is that proposed by Suykens and Vandewalle (1999). $\mathbf{w}$ is the weight vector of the independent variables while b is the intercept. The regularization parameter C scales the error terms u_i^2 and ϕ(X) denotes the kernel function for the feature mapping. Using a quadratic loss function, the model is defined as:

$$\begin{aligned} \min\ J(w, b, u_i) &= \frac{1}{2}||w||^2 + \frac{C}{2}\sum_{i=1}^{N} u_i^2 \\ \text{s.t. } r_i &= w^T\phi(X_i) + b + u_i, \qquad i = 1, ..., N, \end{aligned} \tag{2.10}$$

A solution to the problem can be found by solving the dual form of the Lagrangian function with α_i as the Lagrangian multiplier of the following form:

$$L(w, b, u_i, \alpha_i) = J(w, u_i) - \sum_{i}^{N} \alpha_i(w^T\phi(X_i) + b + u_i - r_i) \tag{2.11}$$

A solution of the dual form based on the Karush-Kuhn-Tucker condition can be obtained with the following linear equation system.

$$\begin{pmatrix} 0 & \mathbf{e^T} \\ \mathbf{e} & \bar{\mathbf{K}} \end{pmatrix} * \begin{pmatrix} b \\ \alpha \end{pmatrix} = \begin{pmatrix} 0 \\ \mathbf{r} \end{pmatrix} \tag{2.12}$$

where $\mathbf{e} = (1, ..., 1)^T$ denotes a $N \times 1$ unit vector, $\mathbf{r} = (r_1, ..., r_N)^T$ is the target vector, $\alpha = (\alpha_1, ..., \alpha_N)^T$ constitutes the Lagrangian multipliers and $\bar{\mathbf{K}} = \mathbf{K} + \frac{1}{C}\mathbf{I}$ with the identity matrix $\mathbf{I}$ and the dimension $N \times N$. So, the final estimated regression model is

$$f(X) = \sum_i \alpha_i^* K(X_i, X) + b^* \tag{2.13}$$

2.3.1.2 Least-squares support vector regression with different intercepts for seniority classes

We enhanced the LS-SVR model to hopefully improve its predictive accuracy in two ways based on Yao et al. (2015). First, we allow for different intercepts b_s for S different seniority classes. So we assume there is some kind of homogeneity within the seniority classes that can be represented by the different intercepts.

$$\begin{aligned} \min\ & J(w, b_s, u_{sj}) = \frac{1}{2}||w||^2 + \frac{1}{2}\sum_{s=1}^{S} b_s^2 + \frac{C}{2}\sum_{s=1}^{S}\sum_{j=1}^{n_s} u_{sj}^2 \\ \text{s.t. } & r_i = w^T\phi(X_{sj}) + b_s + u_{sj}, \qquad j = 1, ..., n_s, s = 1, ..., S \end{aligned} \tag{2.14}$$

The corresponding Lagrangian function results in the following formula.

$$L(w, b_s, u_{sj}, \alpha_{sj}) = J(w, b_s, u_{sj}) - \sum_{s=1}^{S}\sum_{j=1}^{n_s} \alpha_{sj}(w^T\phi(X_{sj}) + b_s + u_{sj} - r_{sj}) \tag{2.15}$$

where $\mathbf{W}$ represents a block diagonal matrix. The dual form emerges in the following optimization problem

$$\min \frac{1}{2}\alpha^T\mathbf{K}\alpha + \frac{1}{2}\alpha^T\mathbf{W}\alpha + \frac{1}{2C}\alpha^T\alpha - y^T\alpha \tag{2.16}$$

2.3.1.3 Semiparametric least-squares support vector regression

We also construct a model which assumes the impact from the different seniority classes is linear. The dummy variables for the seniority classes are z_{sj} and β is is a vector of fixed effects for the seniority of the respective group.

$$\begin{aligned} \min\ & J(w,b,u_i) = \frac{1}{2}||w||^2 + \frac{1}{2}\beta^T\beta + \frac{1}{2}b^2 + \frac{C}{2}\sum_{s=1}^{S}\sum_{j=1}^{n_k} u_{sj}^2 \\ \text{s.t. } & r_i = \mathbf{w}^T\phi(X_i) + \beta^T z_{sj} + b + u_{sj}, \qquad j = 1,...,n_s, s = 1,...,S \end{aligned} \tag{2.17}$$

For this case the Lagrangian function is obtained as

$$L(w,b,u_{sj},\alpha_{sj}) = J(w,b,u_{sj}) - \sum_{s=1}^{S}\sum_{j=1}^{n_s}\alpha_{sj}(w^T\phi(X_{sj}) + b + \beta^T z_{sj} + u_{sj} - r_{sj}) \tag{2.18}$$

Accordingly, with $\mathbf{Z}_{ij} = z_{sj}^T z_{sj}$ and $\mathbf{V}$ as a $N \times N$-matrix of ones, the dual form gives

$$\min \frac{1}{2}\alpha^T\mathbf{K}\alpha + \frac{1}{2}\alpha^T\mathbf{Z}\alpha + \frac{1}{2}\alpha^T\mathbf{V}\alpha + \frac{1}{2C}\alpha^T\alpha - y^T\alpha \tag{2.19}$$

2.3.1.4 ϵ-insensitive support vector regression

Using an ϵ-insensitive loss function the problem is defined in the following form:

$$\begin{aligned} \min_{w,b,u_i,u_i^*} & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N} u_i + C\sum_{i=1}^{N} u_i^* \\ \text{s.t. } & \mathbf{w}^T\phi(x_i) + b - r_i \leq \epsilon + u_i, \\ & r_i - \mathbf{w}^T\phi(x_i) - b \leq \epsilon + u_i^*, \\ & u_i, u_i^* \geq 0, \qquad i = 1,...,N. \end{aligned} \tag{2.20}$$

where $\mathbf{w}$ are the weights of ϵ-SVR, b is the bias, y_i are the targets, C are the costs, N is the number of observations, u_i and u_i^* are the errors, and ϵ is the threshold for tolerated errors.

2.3.2 Regression tree

While SVR has been found by Yao et al. (2015) to exhibit good predictive performance, its explicative capacity is generally low. One class of machine learning methods that has been found to deliver very good predictive performance as well as an easy-to-understand model is the regression tree. Both Qi and Zhao (2011) and Tobback et al. (2014) have used regression trees successfully for LGD modeling. Two other advantages of the regression tree are that it can be used to model non-linearity and it exhibits a relatively robust behavior against outliers. For these reasons, we apply the classification and regression technique (CART) algorithm as defined by Breiman et al. (1984) for the creation of the regression tree model.

2.3.3 Linear regression as benchmark

To investigate whether support vector techniques provide superior predictive ability compared to more traditional approaches, we include a linear regression model as benchmark model. Therefore, we include the following linear regression model for the purpose of comparability as a basic benchmark for the SVR models we propose in this paper:

$$\begin{aligned} r_{ij} =& \alpha + \beta_c(\text{bond characteristics})_{ci} \\ &+\nu(\text{industry distress variables})_{indj} \\ &+\zeta(\text{Principal Components of many macroeconomic variables})_i + \epsilon_{ij}, \\ &\epsilon_{ij} \sim N(0, \sigma^2) \end{aligned} \tag{2.21}$$

2.4 Empirical analysis of recovery rates' prediction

In the following, we explain the dataset that we use in this study and describe the results of our empirical analysis.

2.4.1 Dataset

The initial dataset consists of 794 bonds where, during the period 2002-2012, the issuer filed for a Chapter 11 bankruptcy petition or was assigned a rating by Standard & Poor's of "D" or "SD". This rating agency assigns a rating of "D" if the obligor is in default or in breach of an imputed promise and a rating of "SD" (selective default)

if it believes an obligor rated is in default on one or more of its financial obligations including rated and unrated financial obligations. The recovery rates have been observed sequentially and the panel of observed recovery rates is unbalanced. The bonds have been identified using S&P Capital IQ. All bonds are denominated in US dollars and have a total par value not less than USD 5 million and have no embedded options. The recovery rates are calculated as average volume-weighted prices in the 30 days after the default based on data from TRACE. One macroeconomic variable (number of bonds defaulted) and the industry variables are obtained from Bloomberg, while the macroeconomic variables are found using the database from the Federal Reserve Bank of St. Louis (FRED, Federal Reserve Economic Data).

The companies that issued these bonds can be assigned to the following industries: industry, consumer discretionary, consumer staples, telecommunications, raw materials, utilities, energy, financial services and information technology. We excluded 19 bonds because of one of the following criteria: corrupted data, no matching industry found or a company type other than private or public. As a result, the final sample consists of 775 defaulted bonds.

The average recovery rate for the 775 defaulted bonds in the sample is 40.64% and the sample standard deviation is 29.67%. Recovery rates for companies in an industry that are in distress are much lower than the overall average recovery rate. To avoid having very small seniority classes, the two classes with the fewest observations, junior subordinate and subordinate, are combined. The average recovery rates within the seniority classes comply with the expectation that senior creditors do better: Senior secured bonds have the highest recovery rate of 67.77% followed by the senior unsecured. Accordingly, senior subordinated bonds exhibit the second lowest average recovery and subordinated bonds exhibit the lowest average recovery rate of 7.96%. The frequency distribution across the different seniority classes is shown in Figure 2.2.

2.4.2 Experimental set-up

To decide on the number of principal components, we checked the predictive performance with the number of principal components between 1 and 15 for each variant of PCA. Based on that, we included the first 8 principal components for PCA and

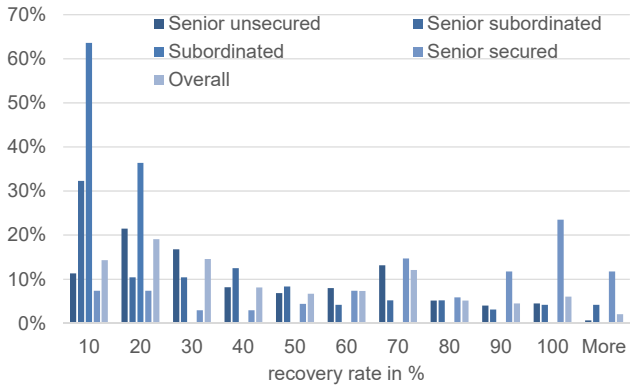

Figure 2.2: Relative frequency distribution of the recovery rates for each seniority class

SPCA. Allowing 20 non-zero loadings for each principal component and using a δ of 0.01, we checked the robustness of our results for SPCA using a grid search for δ and the number of non-zero loadings. The hyperparameters for the autoassociative neural network that we use for NLPCA are chosen by the built-in routine in the neutmatsa toolbox by Hsieh et al. (2006). By using the inconsistency index defined by Hsieh et al. (2007) for cross validation, we reduce the danger of overfitting. For KPCA we follow the suggestion by Wang (2012) and use five times the average distance between one data point and its nearest neighbor as width σ of the radial basis function kernel as defined in Section 2.3.1.

The procedure for choosing the hyperparameters and training the different models is the following. To avoid numerical problems during the LS-SVM computation, the independent variables are scaled to the interval [0,1] according to Hsu et al. (2003). To maintain interpretability of the regression coefficients, we do not scale the variable for the linear regression. For the selection of the model hyperparameters, a ten-folds cross validation (i.e., procedure for assessing the accuracy and validity of a statistical model) is carried out on 70% of the data set. These training data are randomly stratified for their seniority ranks while creating the partitions for the cross validation. The remaining 30% of the data set is used as out-of-sample test set.

During the cross validation, a grid search for the three hyperparameters C, γ and

ϵ of the ϵ-insensitive loss function is carried out in the intervals [2^{-2} to 2^{0}], [2^{-1} to 2^{3}] and [2^{-3} to 2^{3}]. For the quadratic loss function the hyperparameters C and γ are searched for in the intervals [2^{-4} to 2^{5}] and [2^{-4} to 2^{3}]. For each training set partition, the respective model is trained and the performance of the resulting model is evaluated on the respective validation partition. The performance metrics obtained during the ten cross validation parts are averaged for each model configuration to determine the best model's hyperparameters.

As performance metrics, the root mean squared error (RMSE), the coefficient of determination (R^2), the adjusted coefficient of determination (adj. R^2) and the mean absolute error (MAE) are used.

2.4.3 Empirical modeling results

In Table 2.2 we provide an overview of the different model specifications we investigate based on the entire dataset of 775 corporate bonds. The dependent variable is the defaulted bond's recovery rate. The independent variables we use in model (1) are the instrument characteristics. In model (2) we added the macroeconomic variables. We add the first eight principal components obtained from the 104 macroeconomic variables in model (3). In model (4) we add sparse principal components instead of principal components and in model (5) we add nonlinear principal components from an autoassociative neural network. We add the kernel principal components in model (6).

In the linear regression model specification in Table 2.2 the seniority "senior secured" is the omitted dummy variable. Therefore, the negative signs of the seniority dummies and the highly significant coefficients decreasing with lower seniorities are in agreement with the expectation that senior creditors do better than subordinated creditors. This result confirms the observations found in the literature (see, for example, Cantor and Varma (2004)). The dummy variable indicating defaults under Chapter 11 is significantly negative in all three models. This is in accordance with the evidence from Jankowitsch et al. (2014) that recovery rates are lower in formal legal procedures than in informal restructurings.

In Table 2.3, bond characteristics, such as the seniority class, and the industry vari-

Table 2.2: The results of the various regression specifications.

In this table the dependent variable is the recovery rate of the respective bond. The independent variables we have used in model (1) are basic variables of our model. In model (2) we added our five macroeconomic variables. In model (3) we remove the macro factors added in specification (2) and instead, we add the first eight principal components of 104 macroeconomic variables. The respective t-statistics for each factor are presented in parentheses. Statistical significance on the 99% level is indicated with ***, significance on the 95% level is indicated with ** and significance on the 90% level is marked with *.

Variable	Basic(1)	MacroAdded(2)	MacroPCAdded(3)
Intercept	77.1162***	71.48***	66.3534***
	(19.405)	(4.1264)	(17.1303)
UtiInd	11.8588**	10.9264**	2.2108
	(2.1962)	(2.074)	(0.4536)
Financial	7.2244**	9.252***	6.0282*
	(2.1655)	(2.6049)	(1.716)
TelCom	-6.8601*	-4.3032	1.5956
	(-1.8344)	(-1.2071)	(0.3683)
Cons.-Cycl	-6.7359**	-5.7899*	-4.9462*
	(-2.0687)	(-1.9048)	(-1.7098)
Industrial	4.6507	-1.92	-6.1048*
	(1.1272)	(-0.4875)	(-1.6582)
SenUn	-25.0701***	-27.3017***	-26.9766***
	(-6.935)	(-8.1016)	(-8.4175)
SenSub	-31.1246***	-30.5833***	-33.0093***
	(-7.5117)	(-7.9154)	(-9.0752)
Subord	-50.192***	-52.6569***	-51.0351***
	(-5.704)	(-6.4527)	(-6.6178)
TraVol	0.08**	0.06**	0.05*
	(2.492)	(2.0594)	(1.7191)
IndDis1	-13.9111***	-0.8917	8.3057***
	(-5.3438)	(-0.2851)	(2.7442)
IndDis2	-3.1224	3.613	8.5264***
	(-1.2479)	(1.3818)	(3.1228)
VolDef		-0.0152***	
		(-4.0398)	
HYInd		-1.1484**	
		(-2.3144)	
δ **GDP**		-57.7252	
		(-0.7102)	
Unempl.		-1.1122	
		(-0.5173)	
FedFunds		3.6028*	
		(1.8378)	
Chapter 11	-18.6811***	-24.4498***	-22.1121***
	(-8.923)	(-11.1681)	(-10.442)
PC1			-1.9397***
			(-12.9835)
PC2			-0.2686
			(-1.4646)
PC3			3.074***
			(9.114)
PC4			0.657
			(1.4531)
PC5			4.5167***
			(6.9153)
PC6			3.5963***
			(4.9279)
PC7			-0.29
			(-0.2428)
PC8			-2.3131**
			(-2.3172)
Adj. R^2**(in-sample)**	0.2532	0.3597	0.4350
RMSE (in-sample)	25.4280	23.4689	22.0006
MAE (in-sample)	20.4347	18.8939	17.1903
AIC	7.24E+03	7.13E+03	7.03E+03
BIC	7.30E+03	7.21E+03	7.13E+03
# of observations	775	775	775

ables (model (1) of Table 2.2) are included as independent variables. For model (2), the predictive capacity of all models is substantially higher than in the models that do not include the macroeconomic variables. The performance of recovery rate models for defaulted corporate bonds improves by including macroeconomic variables, which is consistent with the findings of previous studies (see, for example Qi and Zhao (2011)).

Table 2.3: Cross validation results of the various models specifications

This table shows the performance measures from cross validation and the respective standard deviations for the models using the independent variables from Table 2.2 for the respective model. The best value for each measure for the respective model is underlined. (LS-SVR: Least Squared Support Vector Regression; LS-SVR DB: Least Squared Support Vector Regression with Different Intercepts; SP LS-SVR: Semi-Parametric Least Squared Support Vector Regression; ϵ-insensitive: ϵ- Support Vector Regression; Lin. Reg.: Linear Regression; Reg. Tree: Regression Tree; model (1): Basic; model (2): MacroAdded; model (3): PCA;

Model (1)	Width	Cost	ϵ	RMSE	σ_{RMSE}	R^2	σ_{R^2}	Adj. R^2	$\sigma_{Adj.R^2}$	MAE	σ_{MAE}
LS-SVR	1	2	-	24.2090	2.2567	0.3074	0.1215	0.2956	0.1236	18.8966	1.6605
LS-SVR DB	1	2	-	<u>24.2038</u>	2.2621	<u>0.3075</u>	0.1228	<u>0.2957</u>	0.1249	18.8734	1.6611
SP LS-SVR	0.125	2	-	24.2094	2.4916	0.3065	0.1372	0.2947	0.1396	<u>18.5942</u>	1.8053
ϵ-insensitive	0.5	1	8	26.6061	1.4174	0.1856	0.0509	0.1717	0.0517	21.4026	0.8100
Lin. Reg.	-	-	-	25.9796	2.1534	0.2076	0.0902	0.1941	0.0917	20.9037	1.5771
Reg. Tree	-	-	-	26.3648	1.8742	0.1918	0.1083	0.1780	0.1102	20.0172	1.5113
Model (2)	Width	Cost	ϵ	RMSE	σ_{RMSE}	R^2	σ_{R^2}	Adj. R^2	$\sigma_{Adj.R^2}$	MAE	σ_{MAE}
LS-SVR	1	2	-	20.2195	2.9789	0.5176	0.1176	0.5061	0.1204	14.2705	1.8932
LS-SVR DB	1	2	-	20.2071	2.9729	0.5182	0.1172	0.5067	0.1200	14.2371	1.9047
SP LS-SVR	0.0625	2	-	<u>18.7271</u>	3.8730	<u>0.5811</u>	0.1556	<u>0.5711</u>	0.1593	<u>12.5029</u>	2.1687
ϵ-insensitive	0.5	1	2	24.7691	1.6096	0.2924	0.0788	0.2756	0.0807	19.8137	0.8192
Lin. Reg.	-	-	-	24.2165	2.5732	0.3072	0.1329	0.2907	0.1361	19.5817	2.0272
Reg. Tree	-	-	-	20.9705	4.0059	0.4738	0.2031	0.4613	0.2079	13.9913	2.3063
Model (3)	Width	Cost	ϵ	RMSE	σ_{RMSE}	R^2	σ_{R^2}	Adj. R^2	$\sigma_{Adj.R^2}$	MAE	σ_{MAE}
LS-SVR	2	2	-	19.4190	3.2283	0.5534	0.1270	0.5410	0.1305	13.1500	1.6915
LS-SVR DB	2	1	-	19.1872	2.9316	0.5647	0.1118	0.5525	0.1149	13.3208	1.6199
SP LS-SVR	0.25	2	-	<u>17.8687</u>	2.5052	<u>0.6271</u>	0.0660	<u>0.6167</u>	0.0679	<u>11.9189</u>	1.2164
ϵ-insensitive	0.5	1	0.5	25.4568	1.6978	0.2559	0.0413	0.2351	0.0425	19.7850	0.9350
Lin. Reg.	-	-	-	22.6437	2.4562	0.3919	0.1283	0.3749	0.1319	17.7387	2.1046
Reg. Tree	-	-	-	21.5982	2.9164	0.4500	0.1486	0.4346	0.1528	14.6812	1.8214

Table 2.4: This table reports the Eigenvalues and the explained variances for the first eight principal components.

PCs	1	2	3	4	5	6	7	8
Eigenvalue	46.9395	29.0396	13.5668	4.9622	2.2578	1.6081	0.9691	0.8570
Cumulative Explained Variance	45.13%	73.06%	86.10%	90.87%	93.04%	94.59%	95.52%	96.35%

For all three performance metrics, the three LS-SVR models show distinctive outperformance compared to the linear regression model for all three performance metrics, a finding that is consistent with Loterman et al. (2012) and Yao et al. (2015) [3]. Among

[3]Because in many cases multiple bonds of the same issuer default at the same time, the draws are not independent. To check our results for robustness, we deleted all bonds of the same issuer with the identical seniority and the identical default date from the dataset. SP LS-SVR still outperforms the other prediction methods and the predictive accuracy is increased when adding the principal components of many macroeconomic variables.

the three tested LS-SVR techniques, the semiparametric LS-SVR yields the best results with an adjusted R^2 of 57.11% which is almost twice the adjusted R^2 of 29.07% of the linear regression model. Both the standard LS-SVR and the LS-SVR with different intercepts for the seniority classes have adjusted R^2-values of 50.61% and 50.67%, respectively, also an impressive outperformance in comparison to the standard linear regression model. However, in assessing performance, it must be recognized that the LS-SVR models also have a slightly higher variance for the performance metrics, indicating that there might be a higher risk of overfitting.

In Figure 2.3 we present the graphics of the first two principal components calculated from the 104 macroeconomic variables shown in Table 2.1 with the procedure described in Section 2.2. The financial crisis years, 2007 and 2008, are shaded in Figure 2.3. Including the principal components of the macroeconomic variables improves the predictive capacity of our models significantly. As can be seen from Figure 2.4, the predictive capacity in the linear regression model increases steadily until the eighth principal component. Moreover, the first eight principal components capture more than 96% of the variance of the underlying dataset. As can be seen in Table 2.4, the first three principal components explain 45%, 28% and 13% of the variance within the macroeconomic variables. Notably, among the 10 variables with the highest communalities in the first 8 principal components are five stock market related variables. Furthermore, the yield on BAA ranked corporate bonds and the inflation expectation by the University of Michigan have the highest and third-highest communalities. As the loadings of the macroeconomic variables are small and it is difficult to identify the most important effects, we apply SPCA to generate principal components that are easier to interpret. The first sparse principal component is most heavily influenced by the number of non-performing loans, the number of bank loans, and the fed funds rate. Real imports, final sales and the PPI of energy goods are the variables with the biggest influence on the second sparse principal component. We find the highest adjusted R^2 by including the first eight principal components. Therefore, we include the first eight principal components from the 104 macroeconomic variables in our analysis in model (3).

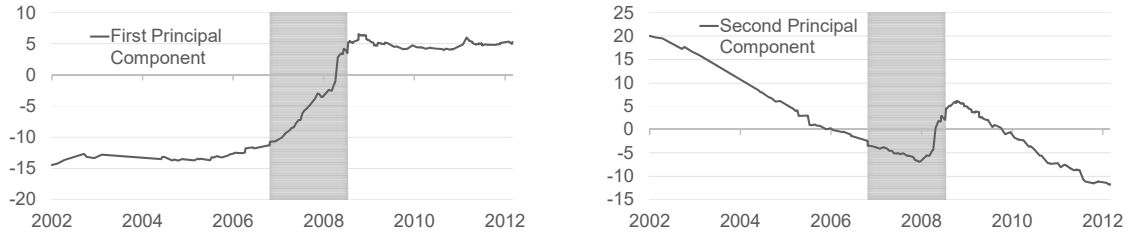

Figure 2.3: The first two principal components from 104 macroeconomic variables

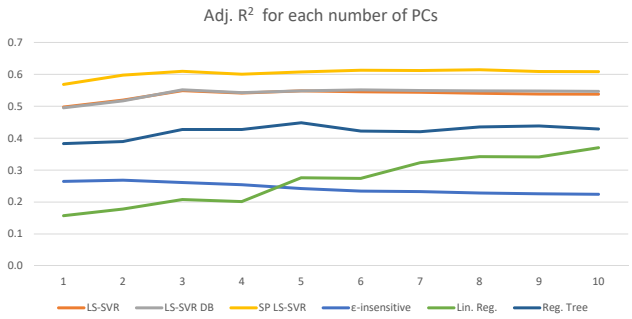

Figure 2.4: Adjusted R^2 for different numbers of principal components in model (3)

For the linear regression the adjusted R^2-value increases to 37.49%, almost twice the value reported for basic model. For the best model, the semiparametric LS-SVR, the adjusted R^2-value is 61.67%. Using the independent variables from model (3), the semiparametric LS-SVR significantly outperforms all the models without the principal components of the 104 macroeconomic variables, while all LS-SVR models also outperform the linear regression significantly.

As can be seen in Table 2.5, applying SPCA and NLPCA for the dimensionality reduction of the 104 macroeconomic variables in models (4) and (5) does not yield a big increase in the out-of-sample prediction quality compared to using PCA. For both compression techniques, semiparametric LS-SVR yields the highest adjusted R^2-value while the other LS-SVR produce the next best results. The regression tree outperforms linear regression and ϵ-insensitive SVR, which shows the least predictive ability. Semiparametric LS-SVR yields a higher adjusted R^2-value of 62.99% for SPCA compared to 61.67% using PCA while it yields an adjusted R^2-value of 57.82% in conjunction with NLPCA. For the other models, making use of SPCA produces similar results to PCA. NLPCA clearly performs worse than PCA using the other weaker models. In particular, for the linear regression the adjusted R^2-value is only 32.13%.

Table 2.5: Cross validation results of the various models specifications

This table reports the performance measures from cross validation and the respective standard deviations for the models using Sparse PCA (model 4), Nonlinear PCA (model 5) and Kernel PCA (model 6) for compressing the 104 macroeconomic variables. The best value for each measure for the respective model is underlined.

Model (4)	Width	Cost	ϵ	RMSE	σ_{RMSE}	R^2	σ_{R^2}	Adj. R^2	$\sigma_{Adj.R^2}$	MAE	σ_{MAE}
LS-SVR	2	4	-	18.9951	3.2835	0.5729	0.1230	0.5610	0.1264	12.7290	1.7750
LS-SVR DB	2	2	-	18.9638	3.0633	0.5752	0.1125	0.5634	0.1157	12.9353	1.6615
SP LS-SVR	0.125	2	-	17.5549	3.2682	0.6400	0.1073	0.6299	0.1103	11.6208	1.9986
ϵ-insensitive	0.5	1	0.5	25.2802	1.7137	0.2661	0.0440	0.2456	0.0452	19.6622	0.9289
Lin. Reg.	-	-	-	22.6129	2.4942	0.3936	0.1292	0.3767	0.1328	17.7487	2.1379
Reg. Tree	-	-	-	21.7079	3.7579	0.4389	0.2118	0.4233	0.2177	14.8254	2.2962
Model (5)	Width	Cost	ϵ	RMSE	σ_{RMSE}	R^2	σ_{R^2}	Adj. R^2	$\sigma_{Adj.R^2}$	MAE	σ_{MAE}
LS-SVR	1	32	-	20.4330	3.2829	0.5034	0.1432	0.4895	0.1472	13.8473	1.9292
LS-SVR DB	1	32	-	20.4325	3.2829	0.5034	0.1432	0.4896	0.1472	13.8465	1.9295
SP LS-SVR	0.0625	2	-	18.4743	4.0089	0.5897	0.1704	0.5782	0.1751	12.3010	2.2192
ϵ-insensitive	8	1	8	26.3330	1.7148	0.2028	0.0573	0.1805	0.0589	20.7718	0.9375
Lin. Reg.	-	-	-	23.5879	2.6100	0.3397	0.1405	0.3213	0.1444	18.9135	2.0685
Reg. Tree	-	-	-	21.8709	1.6671	0.4380	0.1129	0.4223	0.1161	14.5870	1.1100
Model(6)	Width	Cost	ϵ	RMSE	σ_{RMSE}	R^2	σ_{R^2}	Adj. R^2	$\sigma_{Adj.R^2}$	MAE	σ_{MAE}
LS-SVR	2	2	-	19.1139	3.0554	0.5670	0.1209	0.5549	0.1242	12.9988	1.7264
LS-SVR DB	2	2	-	19.0078	3.0530	0.5716	0.1195	0.5597	0.1228	12.8598	1.7052
SP LS-SVR	0.0625	2	-	18.1092	2.4079	0.6159	0.0701	0.6052	0.0720	12.0776	1.2780
ϵ-insensitive	0.5	1	0.25	25.0548	1.7068	0.2788	0.0497	0.2587	0.0511	19.5388	0.9042
Lin. Reg.	-	-	-	22.1141	2.1365	0.4216	0.1107	0.4055	0.1138	17.0671	1.7958
Reg. Tree	-	-	-	21.3519	1.9414	0.4631	0.1202	0.4481	0.1235	14.6925	1.4584

This empirical analysis suggests that by regularizing PCA with the elastic net, as done in SPCA, the predictive accuracy is increased. From a bias-variance tradeoff

perspective, the decrease in variance more than offsets the increase in bias. Moreover, SPCA has several advantages, among which there are computational efficiency, more interpretability in higher-dimension data, high explained variance, and recognition of important variables.

Reducing the dimensionality of the data with KPCA in model (6) as reported in Table 2.5 shows the predictive accuracies that are very similar to the predictive capability using standard PCA. Compared to SPCA, we see a lower adjusted R^2-value of 60.52% for the semiparametric LS-SVR while for linear regression, regression tree, and ϵ-insensitive SVR the adjusted R^2-values are slightly higher using KPCA for dimensionality reduction. In comparison to the other nonlinear dimensionality reduction technique, NLPCA, preprocessing with KPCA yields higher adjusted R^2-values for all prediction methods.

2.4.4 Ranking the macroeconomic variables with gradient boosting and enhanced prediction

Gradient boosting as introduced by Friedman (2001) does not rely on a single strong regression model but on an ensemble of weak-base learners. Each base learner is a regression tree as presented in Section 2.3.4. The target for each additional regression tree is to be maximally correlated to the current negative gradient. Consequently, in each iteration an additional base learner is trained on the error of the ensemble generated so far.

Using the squared error loss function, we apply the LS_Boost algorithm presented by Friedman (2001). Gradient boosting machines are highly robust against outliers, missing values, and inclusion of irrelevant predictors. Moreover, they generate competitive predictions, especially for noisy data. Furthermore, gradient boosting machines allow the ranking of a large number of independent variables by their relative importance. The variables' ranking by gradient boosting can identify new macroeconomic variables not widely used in prior studies. We apply MATLAB's built-in predictorImportance function for the evaluation of each variable's importance. In doing so, MSE in the parent node is compared to the total MSE of the two child nodes. This way of ranking variables also follows the approach to measuring a variable's influence suggested by

Breiman et al. (1984). We use a grid search with a ten-folds cross validation for our choice of the learning rate, the number of trees, and the minimum leaf size of a base learner when building the gradient boosting machine.

From the results reported in Table 2.6, we can gain several insights from the gradient-boosting machine analysis of the macroeconomic variables considered in this paper. There are three groups of independent variables that exhibit significant relative importance. Among the 20 variables with the highest relative importance there are six micro-level factors which are all interest rate related. As can be seen from Table 2.6, the credit spread, the term spread, and the levels of corporate bond and government bond yields have a meaningful impact on the analysis. Moreover, seven stock-market related variables are of particular importance. Four business cycle variables show considerable informativeness. The predictiveness of the variables Housing starts, New orders of capital goods, and the University of Michigan consumer sentiment are characteristic of leading macroeconomic indicators in a prediction task while also normally lagging macroeconomic measure such as the number of unemployed has a noteworthy importance in the gradient boosting machine analysis.

Table 2.6: Relative variable importances (RVIs) of the 104 macroeconomic variables

This table reports the descriptive statistics of the 20 macroeconomic variables that have the biggest relative importance (RVI score) in the gradient boosting framework outlined in section 2.4.4 in the order of their importance.

Variable	Category	Unit	Mean	Std. Dev.	Min	Max
Corporate yield spread	Micro-level factors	Percent	3.523428	1.2846	1.56	6.01
BAA corporate bond yield	Micro-level factors	Percent	6.567655	0.6559	5.78	9.59
AAA corporate bond yield	Micro-level factors	Percent	5.211637	0.5247	3.27	6.58
Russell 2000 return	Stock Market Indicator	Index	-0.09999	0.2116	-0.458	0.5731
Total no unemployed	Business Cycle Indicators	Millions	4.293238	1.3521	2.381	6.635
S&P 500 Volatility	Stock Market Indicator	Percent	0.255854	0.1484	0.0582	0.8481
Treasury bond yield, 10 years	Micro-level factors	Percent	3.375077	0.7813	1.47	5.2
Term Structure	Micro-level factors	Percent	2.278518	1.0414	-0.57	3.82
S&P 500 return	Stock Market Indicator	Index	-0.12578	0.2032	-0.449	0.4323
Wilshire small cap volatility	Stock Market Indicator	Percent	0.319144	0.1714	0.0899	0.873
Nasdaq 100 return	Stock Market Indicator	Index	-0.06714	0.2392	-0.479	0.606
Dow Jones industrial average Vol	Stock Market Indicator	CBOE Vol Index	26.34508	10.456	9.77	68.71
Housing starts	Business Cycle Indicators	Thousands of units	890.884	509.62	478	2273
Uni Michigan consumer sentiment	Business Cycle Indicators	Index 1st Quarter 1966=100	70.28389	9.1778	55.3	97.1
Nasdaq 100 Vol	Stock Market Indicator	Percent	30.14825	10.533	13.79	74.66
Russell 2000 Volatility	Stock Market Indicator	Percent	0.324417	0.1729	0.1059	0.9058
FX index major trading partners	International Competitiveness	Index March (1973=100)	77.67859	4.8613	68.835	105.4
Total borrowings from fed reserve	Financial Conditions	Billions of Dollars	73.45302	77.772	0.01	437.53
New orders: capital goods	Business Cycle Indicators	Billions of Dollars	57.85661	7.0022	46.324	69.04
Number of defaulted bonds	Micro-level factors	Number	750.0503	468.7891	55.0000	1369.0000

The volatilities of all three major US equity market indices and the Wilshire small cap index exhibit significant informativeness according to the ranking of macroeconomic variables. Yet, it does not seem that they have been examined in the recovery

rate literature yet. Housing starts and the University of Michigan consumer sentiment also seem to be two interesting indicators. These highly ranked variables have not been investigated in the literature and as two leading economic indicators they add economic interpretability to any forecasting model. For a number of variables, the ranking from gradient boosting confirms the variable choices of the present literature. Among the most informative variables in our ranking are the corporate yield spread (a variable used in Cantor et al. (2004)), the term structure (a variable used in in Jankowitsch et al. (2014)), and the equity market indices' returns (as used in Altman et al. (2005), Yao et al. (2015) and Cantor et al. (2004)).

Interestingly, while in literature such as in Qi et al. (2011) or Jankowitsch et al. (2014) usually the short end of the interest rate curve (rank 83) is examined, in our analysis we find that the 10-year Treasury yield among the most informative variables. Several popular variable choices such as GDP growth (rank 61), as for example included by Altman et al. (2005), and the unemployment rate (rank 76) , for example included in Yao et al. (2015), are not among the 20 most informative variables. However, the number of unemployed, which is highly correlated to the unemployment rate, is among the most informative variables. We use the number of defaulted bonds in the year before the default as a proxy for the default rate, which is often included in the literature such as in Qi et al. (2011). The number of defaulted bonds exists in the list of informative macroeconomic variables.

In model (7) we added the 20 variables that have been identified as most informative by gradient boosting to the bond data from our base model (1). The results are reported in Table 2.7. This modification yields the highest R^2-values for the LS-SVR (56.70%), LS-SVR with different biases (56.75%), the regression tree (48.10%), and the linear regression (42.82%). The predictive performance for semiparametric LS-SVR is generally close to the accuracies of the PCA techniques, only lagging behind semiparametric LS-SVR when SPCA is employed. Selecting macroeconomic variables by gradient boosting is useful because it significantly enhances the predictive accuracies of models that are easy to interpret, such as the regression tree and the linear regression.

Table 2.7: Cross validation results of the model selecting macroeconomic variables. The best value of each measure for the respective model is underlined.

This table reports the performance measures from cross validation and the respective standard deviations for the models using gradient boosting to select the 20 most informative macroeconomic variables.

Model (7)	Width	Cost	ϵ	RMSE	σ_{RMSE}	R^2	σ_{R^2}	Adj. R^2	$\sigma_{Adj.R^2}$	MAE	σ_{MAE}
LS-SVR	1	4	-	18.7160	3.3805	0.5855	0.1259	0.5670	0.1315	12.1433	1.7244
LS-SVR DB	1	4	-	18.7064	3.3718	0.5860	0.1253	0.5675	0.1308	12.1227	1.7161
SP LS-SVR	0.0625	2	-	<u>17.8390</u>	2.5920	<u>0.6291</u>	0.0628	<u>0.6126</u>	0.0655	<u>11.9428</u>	1.3672
ϵ-insensitive	0.5	1	2	24.3329	1.9425	0.3164	0.0967	0.2860	0.1010	19.1144	1.0587
Lin. Reg.	-	-	-	21.5054	1.9363	0.4526	0.0979	0.4282	0.1023	16.4414	1.6507
Reg. Tree	-	-	-	20.4652	2.9565	0.5031	0.1588	0.4810	0.1659	13.7896	1.8602

2.5 Conclusions

The recovery rate is a key parameter in the Basel II/III accords and it is one of the main risk factors in pricing financial products and contracts related to credit risk. In this paper, we study the performance gain by using support vector techniques, linear regression, and regression tree for predicting recovery rates of defaulted corporate bonds with different explanatory variables. A LS-SVR model with different biases for the different seniority classes exhibited very similar predictive performance to the standard LS-SVR. A semiparametric LS-SVR model which takes the seniority indicator dummies as linear input showed significant outperformance, not only versus the linear regression model, but also in comparison to the standard LS-SVR approach. We find that the SVR approaches outperform the linear regression in terms of out-of-sample adjusted R^2, RMSE, and MAE. The recovery rate literature has applied models that can only handle a limited number of independent variables.

This paper contributes to the literature on corporate bond recovery rate prediction in four ways. First, in contrast to the literature, which used a few macroeconomic variables in predicting recovery rate, the empirical evidence confirmed that by adding the principal components derived from 104 macroeconomic measures from a broad range of categories such as stock market conditions, lending conditions, international competitiveness, business cycle conditions, and micro-level conditions the predictive capacity of our models increased. Second, using SPCA instead of PCA, the predictive capacity of the models increases, while the results using NLPCA and KPCA instead of PCA are mixed. Furthermore, SPCA allows a better interpretability of the principal components. Third, we ranked a large set of macroeconomic variables by gradient boosting, from best to worst, based on their overall predictive power of recovery rate. The analysis indicates that the most informative macroeconomic variables in predicting

recovery rates of U.S. corporate bonds are the credit spread of corporate bonds, the yields offered on corporate bonds, the annual return of the Russell 2000, and the number of unemployed. We introduced new macroeconomic variables not commonly used in prior research, for example housing starts, orders of capital goods, and stock market volatility. Fourth, adding the 20 most informative variables to the base model increases the predictive accuracy of models, that are easy to interpret, such as the regression tree and the linear regression as well as three of the four types of LS-SVR. Overall the empirical results of this study show that including a large number of macroeconomic variables yields better estimates of the recovery rate.

Chapter 3

Fuzzy decision fusion approach for loss-given-default modeling

This chapter is joint work with Dr. Abdolreza Nazemi[1], Farnoosh Fatemi Pour[2], and Prof. Frank J. Fabozzi[3] published in 2017 as: Fuzzy decision fusion approach for loss-given-default modeling, European Journal of Operational Research, 262(2), 780-791. https://doi.org/10.1016/j.ejor.2017.04.008

3.1 Introduction

According to Basel II and Basel III Accords, banks in the G20 countries need to hold capital requirements for managing their risk based on expected loss. There are three key parameters in the Internal Rating Based (IRB) advanced approach for calculation of expected loss. These parameters are probability of default, loss-given-default (LGD), and exposure at default. Consequently after the Basel II accord, LGD has become a much more vital measure for financial institutions, especially banks. Therefore, financial institutions need reliable LGD predictions. In this paper, we propose a new method for prediction of LGD based on fuzzy rule-based models and ensemble methods incorporating a broad range of macroeconomic variables that provide a significant increase in performance measures compared to methodologies proposed in the literature. In our proposed model, a fuzzy rule-based model is trained using a differential evolution algorithm to dynamically weight ensemble members and combine their decisions to obtain the final output. The differential evolution algorithm is used

[1] School of Economics and Business Engineering, Karlsruhe Institute of Technology
[2] Departments of Computer and Electrical Engineering, Ferdowsi University of Mashhad, Iran
[3] EDHEC Business School, Nice, France

to be able to better cope with complex data and avoid the curse of dimensionality in creating fuzzy rule-based models.

Using multiple models to analyze the data provide different insights about the same data. Decision fusion is the combination of multiple decision makers. Fusing the decisions made by multiple different decision makers enables one to benefit from multiple views instead of one. For this reason, we fuse the results of multiple techniques including a linear regression model, four types of support vector regression (SVR) techniques and a decision tree in our fuzzy rule-based model.

In this paper, we make three contributions to the literature: First, fuzzy decision fusion models are applied to modeling LGD of corporate bonds for the first time. Second, the technique we propose improves predictive accuracy of our models by adding the principal components derived from 104 macroeconomic variables. Finally, in order to improve the predictive accuracy of the models, Box-Cox transformation of macroeconomic variables is tested. We show that fuzzy decision fusion models outperform all types of SVR techniques, decision trees and regression in both terms of prediction accuracy and interpretability.[4]

The remainder of the paper is organized as followed. A review of the techniques used in the literature to predict LGD on debt obligations is provided in Section 3.2. In Section 3.3, we present the models we propose and in Section 3.4 we describe the data used. We then present the empirical results in Section 3.5. Section 3.6 concludes the paper.

3.2 Literature review

Two types of predictive models have been applied in the literature: parametric and non-parametric. Among the parametric models, the most popular are linear regression models that have shown robustness and effectiveness in LGD prediction and explanation. Zhang and Thomas (2012) report poor out-of-sample performance for linear

[4] As Gacto et al. (2011) note, interpretability is the capacity to express the behavior of the real system in an understandable way. In the approach we propose, a fuzzy rule-based model is used for weighting the base models described in Section 3.3.

regression and survival regression for modeling LGD of personal loans. Calabrese and Zenga (2010) propose a beta regression model to predict recovery rates of loans. Leow and Mues (2012) apply a two-stage model with a combination of a probability of repossession model and a haircut model for the LGD of residential mortgage loans. Jacobs and Karagozoglu (2011) apply a beta-link generalized linear model to predict LGD. Loterman et al. (2012) compare the predictive accuracy of 24 techniques for the prediction of LGD in different datasets. They find that the predictive accuracy of non-parametric techniques such as support vector machines and artificial neural networks are higher than the typical linear regression models. Bellotti and Crook (2012) build several models for retail credit cards LGD such as the tobit model, decision tree model and ordinary least squares (OLS) model. Parametric methods had weak results for LGD modeling in their study. The main advantage of parametric models is their interpretability, but they usually have weak prediction power in LGD modeling.

Hartmann-Wendels et al. (2014) predicted the LGD for defaulted leasing contracts from three German leasing companies. They reported that out-of-sample LGDs estimation is necessary for appropriate risk management. In their large sample, model trees outperformed other methods. Yang and Tkachenko (2012) applied different parametric and non-parametric techniques for modeling exposure at default and LGD for commercial borrowers. Bastos (2010) applied a regression tree and parametric fractional response regression for modeling LGD of bank loans. He empirically demonstrates the better prediction capability of regression trees using a data set consisting of 374 defaulted loans. Gürtler and Hibbeln (2013) introduced several improvements for LGD of defaulted bank loans. Park and Bang (2014) considered different factors of defaulted mortgages such as borrower characteristics, foreclosure auction process, seniority, housing type, housing market cycle and submarkets for the LGD modeling of residential mortgages in Korea. Park and Bang (2014) report that the recovery rate mean for senior mortgages is significantly higher than the subordinated claims and show the effects of housing market cycles on LGD.

Altman et al. (2005) studied the relation between LGD and aggregate default rates on corporate bonds from 1982 to 2002. They find that LGDs of corporate bonds are related to default rates, seniority and collateral levels. Moreover, they report the corporate bond market variables explain more variation in the LGD than macroeconomic factors. Cantor and Varma (2004) show that seniority, security, industry and macroe-

conomic factors are correlated to LGD. Analyzing 3,751 US corporate bonds and loans for the period 1985-2008, Qi and Zhao (2011) report that neural networks outperform parametric models. Altman and Kalotay (2014) propose an approach based on the mixtures of Gaussian distributions to forecasting the distribution of ultimate recoveries on defaulted loans and bonds. Their method outperforms parametric regressions as well as regression trees as a non-parametric benchmark. They present more evidence of industry-driven effects on the forecasting of the distribution of LGD.

Yao et al. (2015) predict LGD for corporate bonds employing three different types of SVR techniques and parametric methods. They report the performance of SVR is significantly higher than parametric techniques such as a fractional response regression or a multiple linear regression. Applying parametric and non-parametric methods for predicting LGD for corporate bonds, Nazemi et al. (2018a) mention that SVR techniques exhibit significantly higher predictive performance than a regression model and decision tree techniques.

Duffie et al. (2009) and Koopman et al. (2011) report the influences of macroeconomic variables for the probability of default calculation. Cantor and Varma (2004) report that macroeconomic variables are important variables in LGD estimation. Analyzing UK data for major retail credit cards, Bellotti and Crook (2012) find that OLS models that include macroeconomic variables have the best performance for estimating LGD. Investigating the importance of macroeconomic independent variables in retail loans LGD datasets, Leow et al. (2014) find that macroeconomic variables are able to improve the performance of models. Tobback et al. (2014) report that the inclusion of macroeconomic variables can improve the prediction of LGD for corporate loans and revolving credit. Bruche and Gonzalez (2010) state that firms that defaulted during recessions recover less and the number of defaulting firms rises in this period. Chen (2010) mentions that LGD during the recessions of 1982, 1990, 2001, and 2008 are more than average. Nazemi et al. (2018a) find that by incorporating the principal components derived from macroeconomic variables, the predictive performance of all SVR techniques, as well as the linear regression, increases significantly.

Significantly lower recovery rates for defaulted firms if the industry of defaulted firms is in distress were reported by Acharya et al. (2007). However, Mora (2015) shows that (1) macroeconomic variables are important factors in LGD for corporate

bonds, (2) industries which are more dependent on the global and national economies have lower recovery rates when the stock market drops, and industries whose sales are more related to GDP growth recover less during macroeconomic downturns.

Both industry works and academic papers (such as Loterman et al. (2012), Qi and Zhao (2011)) sometimes use the non-parametric regression in LGD modeling. As of this writing, there is no industry practice and academic research that has implemented a fuzzy fusion approach for LGD modeling. Moreover, previous studies have not considered many macroeconomic variables in LGD analysis. This paper aims to fill these gaps in LGD modeling.

Table 3.1 summarizes studies on LGD modeling for corporate bonds.

Our LGD modeling for corporate bonds is close to that proposed by Yao et al. (2015), who investigate SVR techniques and 13 other algorithms for LGD modeling incorporating four macroeconomic variables. Studying the LGD of US corporate bonds for the period from 1985 to 2012, they reported that SVR techniques significantly outperform other methods. Moreover, logistic and beta transformations of LGD do not improve prediction accuracy. This paper has three main contributions compared to Yao et al. (2015) and the existing literature on LGD. First, fuzzy decision fusion models are applied to modeling LGD for the first time. However, we show that fuzzy decision fusion models have significantly higher predictive accuracy compared to all types of SVR models used by Yao et al. (2015). Second, we propose improving predictive accuracy of our LGD models by adding the principal components derived from 104 macroeconomic variables. Finally, we apply the Box-Cox transformation to the macroeconomic variables in order to improve the predictive accuracy of our model.

3.3 Fusion-based loss-given-default modeling

This section details our proposed approach. The proposed approach uses different types of regressors including SVR, regression tree, and OLS regression as base models and trains a fuzzy rule base to combine their results effectively in order to make a prediction of LGD as reported in Section 5. We begin by describing the base models. Then the proposed fuzzy rule-based model is described in detail.

Table 3.1: Overview of models in literature, focusing on US corporate bond.

Acharya et al. (2007)	
Data	1982-1999, Various debt instruments of 300 issuers from S&P/PMD database
Models	Linear regression
Highest R-square	0.68
Main finding	Industry conditions at the time of default are important determinants of creditor recoveries
Altman & Kalotay (2014)	
Data	1987-2011, 4720 debt instruments (which 60% are bonds)
Models	Mixture of Gaussian
Highest R-square	-
Main finding	Mixtures of Gaussian outperform
Bastos (2014)	
Data	1987-2010, 4630 loans and bonds from Moody's URD
Models	Parametric regressions, regression tree
Highest R-square	-
Main finding	Fractional response regression outperform (long horizons), regression tree (short horizons)
Cantor & Varma (2004)	
Data	1983-2003, about 1100 issuers for both loans and bonds
Models	Regression
Highest R-square	-
Main finding	Specify important factors that impact recovery rates
Jacobs & Karagozoglu (2011)	
Data	1985-2008, Corporate loans and bonds
Models	Beta-link generalized linear model
Highest R-square	0.6119
Main finding	Determine important factors in ultimate LGD
Jankowitsch et al. (2014)	
Data	2002-2010, 2235 event/bond combinations
Models	Linear Regression
Highest R-square	0.66
Main finding	Determine important factors in LGD modeling
Mora (2015)	
Data	1970-2008, 4422 instruments
Models	Regression
Highest R-square	0.546
Main finding	The macroeconomic factors act differentially at the industry level
Qi & Zhao (2011)	
Data	1985-2008, 3751 loans and bonds from Moody's URD
Models	RT, NN, fractional response regression, Inverse Gaussian Regression
Highest R-square	0.576
Main finding	Non-parametric methods outperform
Renault & Scaillet (2004)	
Data	1981-1999, 623 bonds from S&P / PMD database)
Models	Kernel estimation, Nonparametric, Monte Carlo
Highest R-square	-
Main finding	Important factors for LGD, recovery rates are far from being beta distributed
Rösch & Scheule (2014)	
Data	1982-2009, 1,653 bonds default and recovery events
Models	Tobit
Highest R-square	-
Main finding	Bond ratings, bond issue characteristics, bond issuer characteristics and macroeconomic variable explain default probabilities and LGDs
Yao et al. (2015)	
Data	1985-2012, 1413 bonds from Moody's URD
Models	Regression, Fractional Response Regression, Support Vector Regression, Two-stage Model
Highest R-square	0.7006
Main finding	LS-SVR outperform, standard transformations of LGD do not improve prediction accuracy

.

3.3.1 Support vector regression

Nazemi et al. (2018a) and Yao et al. (2015) have attempted to predict corporate bond LGD using SVR techniques. Chalup and Mitschele (2008) argue that such techniques are promising for finance applications because of their ability to deal with nonlinear input data. Consequently, least-squares SVR (LS-SVR) as a "kernelized" version of the traditional linear regression is likely to yield a higher predictive capacity.

As demonstrated by Aizerman et al. (1996), Mercer's theorem enables a computationally efficient calculation of a kernelized problem. Therefore, an appropriate kernel function has to be chosen. The only prerequisites a kernel has to fulfill are to be positive semi-definite and to represent a similarity measure between pairs of input samples, as explained by Chalup and Mitschele (2008). In particular, in all SVR models we use the radial basis function kernel in the following form:

$$K(x_i, x_j) = \exp\left(-\frac{||x_i - x_j||^2}{2\sigma^2}\right) \tag{3.1}$$

3.3.1.1 Least-squares support vector regression

In the following, we make use of three different LS-SVR models that we implement in MATLAB. At first, we use the original LS-SVR set-up as proposed by Suykens and Vandewalle (1999), where ω is the feature vector while b denotes the intercept. The regularization parameter C normalizes the error terms u_i^2 and $\phi(x)$ is the kernel function for mapping the features into the higher dimensional space as defined in equation (3.1). N is the number of defaulted bonds while i is the number of the respective bond. The model with a quadratic loss function is defined as:

$$\begin{aligned} \min\ & J(w, b, u_i) = \frac{1}{2}||w||^2 + \frac{C}{2}\sum_{i=1}^{n} u_i^2 \\ \text{s.t. } & lgd_i = w^T\phi(X_i) + b + u_i, \qquad i = 1, ..., N \end{aligned} \tag{3.2}$$

where X_i is explanatory variables for the LGD model that they consist of bond characteristics, industry distress variables and principal components of macroeconomic variables.

From Yao et al. (2015) we build two modifications of a LS-SVR. We model K different seniority classes with different intercepts b_k. We suppose there is some commonality within the seniority classes that can be modeled by using different intercepts.

$$\begin{aligned} &\min\ J(w, b_k, u_{kj}) = \frac{1}{2}||w||^2 + \frac{1}{2}\sum_{k=1}^{K} b_k^2 + \frac{C}{2}\sum_{k=1}^{K}\sum_{j=1}^{n_k} u_{kj}^2 \\ &\text{s.t. } lgd_i = w^T\phi(X_{kj}) + b_k + u_{kj}, \qquad j = 1, ..., n_k, k = 1, ..., K \end{aligned} \tag{3.3}$$

We also build a model wherein we map the influence of the different seniority classes to be linear. z_{kj} denotes the dummy variables for the seniority classes and β denotes the fixed effect for each seniority class.

$$\begin{aligned} &\min\ J(w, b, u_i) = \frac{1}{2}||w||^2 + \frac{1}{2}\beta^T\beta + \frac{1}{2}b^2 + \frac{C}{2}\sum_{k=1}^{K}\sum_{j=1}^{n_k} u_{kj}^2 \\ &\text{s.t. } lgd_i = w^T\phi(X_{kj}) + \beta^T z_{kj} + b + u_{kj}, \qquad j = 1, ..., n_k, k = 1, ..., K \end{aligned} \tag{3.4}$$

where ω is the parameter vector of the associated explanatory variables for the LGD model. When using an ϵ-insensitive loss function our optimization problem becomes the following:

$$\begin{aligned} \min_{w,b,u_i,u_i^*} &\frac{1}{2}w^T w + C\sum_{i=1}^{N} u_i + C\sum_{i=1}^{N} u_i^* \\ \text{s.t. } &w^T\phi(X_i) + b - lgd_i \leq \epsilon + u_i, \\ &lgd_i - w^T\phi(X_i) - b \leq \epsilon + u_i^*, \\ &u_i, u_i^* \geq 0, \qquad i = 1, ..., N. \end{aligned} \tag{3.5}$$

where ω denotes the weights of ϵ-SVR, b is the intercept, lgd_i denotes the LGD, C is the regularization parameter, u_i and u_i^* are the predicted errors, and ϵ is the threshold for tolerated errors.

3.3.2 Regression tree

The basic idea of tree construction is to find subsets with maximum homogeneity or cases that are located in a subset belonging only to one class of a target variable. At each splitting step, tree algorithms split cases with independent variables that have maximum homogeneity. The reduction of impurity that the split obtains is defined as the quality of a split as:

$$\triangle i = i(v) - [\pi(l)i(v_l) + \pi(r)i(v_r)] \tag{3.6}$$

where $i()$ is impurity, $\pi(l)$ and $\pi(r)$ denote the proportions of observations that are sent to the left child node (v_l) or right child node (v_r). In fact, tree algorithms select the variable that allows for the best quality of a split. Finally, classification trees label leaf nodes corresponding to the majority class of target variables. Regression trees fit $\hat{y}_i$ equal to the mean value of the dependent variable of observations at the leaf. The main advantages of the regression tree are that they are (1) easy to understand, (2) exhibit relatively robust behavior against outliers, and (3) are capable of modeling nonlinearity.

3.3.3 Linear regression as a benchmark

We apply a linear regression model to analyze the LGD of US corporate bonds.[5] Two different model configurations regarding the covariates are tested. In the first model the instrument specific variables, which are the seniority and the trading volume, are taken into account. In this model, industry specific variables indicating whether the respective industry was in a state of distress in the year preceding the default are also used. In another model we add the principal components calculated from the 104 macroeconomic variables capturing 96% of the variance within the original macroeconomic data. We propose the following linear regression model:

$$\begin{aligned} \text{loss-given-default} =& \alpha + \zeta(\text{bond characteristics}) + \gamma(\text{industry distress variables}) \\ &+ \psi(\text{Principal Components of many macroeconomic variables}) + \epsilon, \\ &\epsilon \sim N(0, \sigma^2) \end{aligned} \tag{3.7}$$

3.3.4 Fuzzy decision fusion

Assume we have k different sources which are trained using d training datasets which are not necessarily independent. Suppose $h_i(x)$ is the output of decision maker i about the incoming data x, in which $i = 1, \ldots, k$. A decision-fusion based model uses a function, $g(x)$, to combine the results of the base models and the output is used to make the final decision. Therefore, the decision about the input data x is made in the

[5] We use the built-in implementations from MATLAB for both linear regression and the regression tree.

following form:

$$\mathrm{F(x)=g(H(x))} \tag{3.8}$$

In this function, $H(x)$ is the vector containing the outputs of the base models. $H(x) = [h_1(x), h_2(x), \ldots, h_k(x)]$

Decision fusion-based models seek to find the function g which suitably combines the base models. One of the most often used combination functions found in the literature is the following weighted linear combination

$$\mathrm{F(x)=}\sum_{i=1}^{k} W_i h_i(x) \tag{3.9}$$

In this formula, W_i is the weight assigned to the i^{th} model. In the linear combination, we aim to find the best weights for the models. This could be applied at either the training phase (static methods (Burduk and Walkowiak, 2015)) or the execution phase (dynamic methods (Britto Jr et al., 2014; Dos Santos et al., 2008)). Dynamic methods have shown higher accuracy in the literature due to their focus on the input data for weighting models (Jurek et al., 2014).

Fuzzy decision fusion (Loskiewicz-Buczak et al. (1994)) can be applied as a combiner for the fusion of the outputs of base models. One of the main benefits of using fuzzy logic is its ability to avoid crisp boundaries and its power in handling uncertainty. In this paper we propose to train a fuzzy rule base which dynamically weights base models regarding the data being analyzed.

Two phases are typically involved in the generation of fuzzy rules from numerical data: how to partition a pattern space into fuzzy subspaces and how to define a fuzzy rule for each fuzzy subspace (Nozaki et al. (1996)). First, the pattern space is partitioned into some fuzzy subspaces. After this, one or more rules are defined for each subspace created. Each rule makes a decision about any data taking into consideration only the corresponding subspace. For making the final decision, the outputs of all rules are suitably combined. One of the basic methods for partitioning the input space is the simple fuzzy grid method (Ishibuchi et al. (1992)). This method suffers from the disadvantage that the performance of the final rule base directly depends on the chosen parameters for the partition sizes. Another disadvantage of this method is that the number of generated rules might be enormous, especially in the case of

high-dimensional and complex data.

In this paper, we implement the differential evolution (DE) algorithm in MATLAB to create the fuzzy rule base in order to avoid the curse of dimensionality for fuzzy rule-based models and generate an appropriate number of rules forming a high performing rule base. DE, an optimization algorithm proposed by Storn et al. (1997), iteratively optimizes a candidate solution seeking to find the globally optimized solution. In the proposed methods, the same defuzzification formula which is selected for making the final decision of the fuzzy rule base is used as the quality measure in the DE algorithm. The partitioning of the input space is fully done by DE optimization algorithm. We fix the number of generated rules in the DE algorithm.

The final constructed rule base contains R rules in the following form:

$$\begin{aligned} &\text{if x is } \mathrm{C}_1 \; then \;\; [w_1^1, w_2^1, \ldots, w_k^1] \\ &\text{if x is } \mathrm{C}_2 \; then \;\; [w_1^2, w_2^2, \ldots, w_k^2] \\ &\qquad\qquad\qquad\qquad\qquad\qquad \ldots \\ &\text{if x is } \mathrm{C}_R \; then \;\; \left[w_1^R, w_2^R, \ldots, w_k^R\right] \end{aligned} \tag{3.10}$$

where k is number of sources and C_r, $r = 1, \ldots, R$ is the center for r^{th} rule. $[w_1^r, w_2^r, \ldots, w_k^r]$ is the vector of weights that the r^{th} rule assigns to the base models. Figure 3.1 shows the steps for creating the fuzzy rule base.

When unseen data arrive and must be analyzed, its membership value to each rule is calculated. Membership value of each data to each rule is calculated as the inverse of the distance from the center of the rule indicated by the antecedent part of the rule. Several defuzzification formulas can be used in order to get the final results from the fuzzy rule base. We use four different types of defuzzification formula: maximum formula, maximum of maximums, mean formula, and mean of maximum formulas. Below we describe each.

Maximum formula: A rule with maximum membership value is selected and the vector of weights provided by the selected rule is used for fusing the results of the base models as:

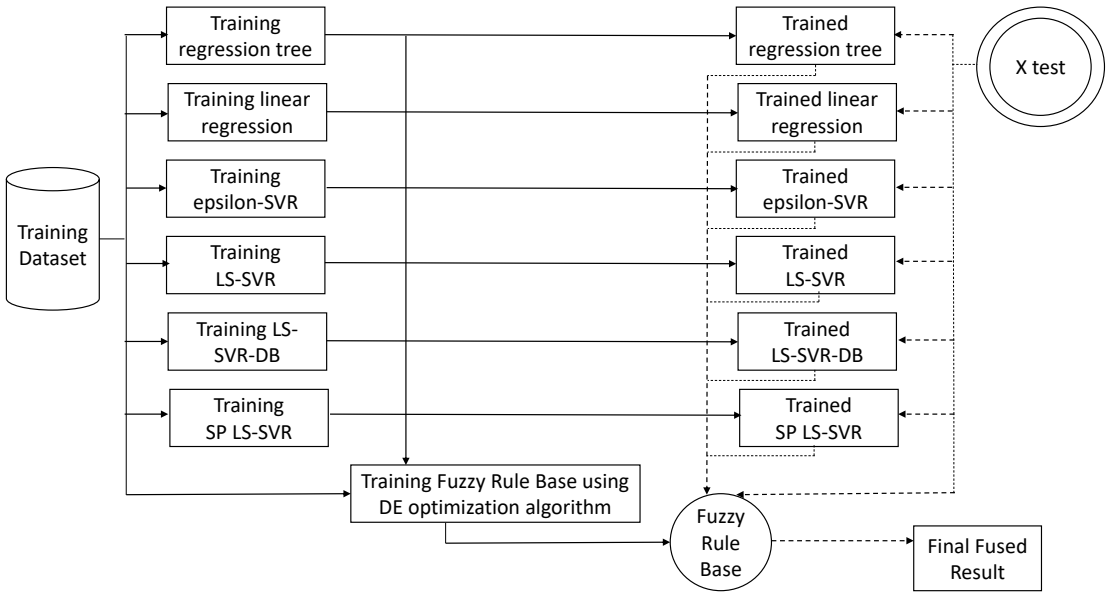

Figure 3.1: The steps for creating fuzzy rule base

$$F(x) = \frac{\omega^{r_{max}} * H(x)}{\sum_{i=1}^{k} \omega_i^{r_{max}}} \tag{3.11}$$

where $H(x) = [H_1(x),\, H_2(x), \ldots,\, H_k(x)]$.

Maximum of maximums: A rule with maximum membership value is selected to provide the final output for the data in this type of defuzzification. The output of the rule is the result of the source with maximum weight as:

$$\mathrm{F(x)}{=}\mathrm{H}_{argmax\,(\omega_{r_{max}})}\ (x) \tag{3.12}$$

Mean formula: With this type of defuzzification, the mean result of different rules for each data is selected as the final decision in the following form:

$$F(x) = \frac{\sum_{i=1}^{R} \omega^{r_i} * H(x)}{k * R} \tag{3.13}$$

Mean of maximums formula: The mean result of different rules for each data is selected as the final decision. Each rule provides its output as the result of base model

with maximum weight as:

$$F(x) = \frac{\sum H_{argmax\,(\omega_{r_{max}})}\,(x)}{R} \tag{3.14}$$

In this way we dynamically weight base models and benefit from fusion of all base models instead of selecting only one.

3.4 Data

In our initial dataset we obtained 794 U.S. default events from the Standard & Poor's Capital IQ database that occurred from 2002 to 2012. Then we filtered for events when the issuer filed for a bankruptcy under either Chapter 7 or Chapter 11 under the U.S. bankruptcy code. Moreover, we included bonds of issuers that Standard & Poor's had assigned a rating of 'D' (default) or 'SD' (selective default). We restrict our analysis to straight bonds denominated in US dollars and have a face value of at least USD 5 million. Based on TRACE data the respective recovery rates are calculated as volume-weighted average trading prices in the 30 days following the default. Our macroeconomic variables are obtained from the Federal Reserve Bank of St. Louis (FRED, Federal Reserve Economic Data). The number of defaulted bonds, industry and stock variables are retrieved from Bloomberg.

The issuers of the bonds in our dataset operate in the following wide range of industries: utilities, energy, financial services, information technology, industry, consumer discretionary, consumer staples, raw materials and telecommunications. Eighteen bonds were excluded from our dataset for one of the following reasons: the data were corrupt, no industry could be retrieved or the company type was neither private nor public. As a consequence, 776 defaulted bonds remained for our analysis. Table 3.2 presents the frequency and the descriptive statistics for each industry. The frequency and summarized statistics of all seniorities are listed in Table 3.3. The frequency and density of recovery rates for all observations are shown in Figure 3.2.

Table 3.2: Descriptive Statistics of recovery rates across industry characteristics

Industry	# of defaults	# of firms	Median	Mean	Std. Dev.
Utilities	31	7	63,37	64,15	24,18
Financials	269	18	25,92	36,55	25,87
Materials	43	16	35,89	41,51	27,51
Communications	120	23	22,42	38,35	35,59
Consumer, cycl.	190	58	31,14	38,34	26,73
Consumer, non-cycl.	27	17	27,56	36,36	30,38
Energy	22	12	45,05	53,09	25,92
Technology	8	3	45,67	51,05	34,15
Industrial	66	28	55,51	52,12	36,34
Overall	776	128	32,97	40,58	29,67

Table 3.3: Descriptive Statistics of recovery rates for each seniority

	Subord	SenSub	SenUn	SenSec	Total
Mean RR	7.96%	34.23%	39.11%	67.77%	40.58%
σ	6.58%	31.97%	27.34%	32.37%	29.69%
$q_0.1$	1.78%	2.14%	9.42%	10.32%	5.91%
$q_0.25$	2.06%	7.08%	16.41%	49.48%	15.51%
Median RR	4.32%	24.13%	30.59%	74.03%	32.97%
$q_0.75$	15.61%	53.67%	62.69%	94.00%	65.58%
$q_0.9$	17.22%	84.09%	77.58%	101.29%	84.68%
# of Values	11	96	601	68	776
% of Values	1.42%	12.37%	77.45%	8.76%	100%

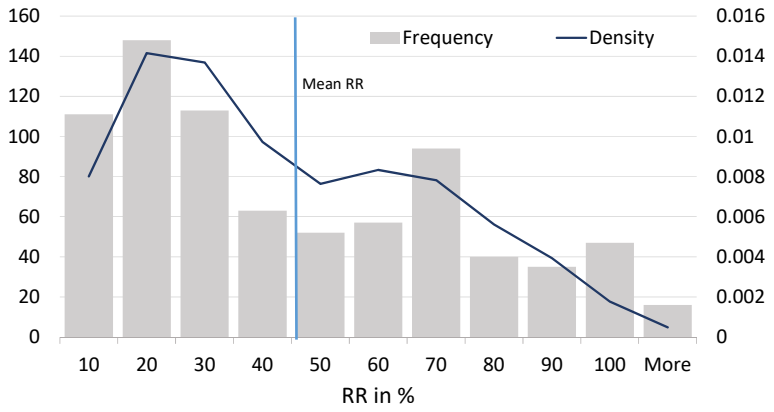

Figure 3.2: The frequency and density of recovery rates (RR)

3.5 Experimental results

We apply a stratified sampling strategy in order to have the same ratios of different bond seniorities in each split of the cross validation. A five folds cross validation is executed for the selection of the SVR hyperparameters. Following Hsu et al. (2003), we scaled explanatory variables to the interval [0,1] for preventing computation problems in the LS-SVM fitting. To ensure the robustness of the results, the performance metrics are obtained from the five folds cross validations. The cross validation performance metrics, the root mean squared error (RMSE), the adjusted coefficient of determination (Adj. R^2) and the mean absolute error (MAE) on the testing sets are reported.

The included variables have been chosen based on a review of the literature. For the basic model (1) the explanatory variables X_i are the bond characteristics and dummies indicating industry distress. In particular, we include the seniority class of the bond in the capital structure, the amount of the bond's trading volume and dummy variables for each of the following sectors: utility, financial, telecommunication, consumer cyclical and industrial. Moreover, we include two variables measuring whether the respective industry is in distress. One industry distress variable indicates whether

industry's sales growth in the year before the default was negative. The other industry distress variables indicates whether the performance of the industry index in the year before the default was worse than -30%.

We include the extensive range of macroeconomic factors such as credit market factors, stock market indices, international competitiveness, financial conditions, business cycle conditions and micro-level factors shown in Table 4. We add the principal components derived from these macroeconomic measures to our model. Moreover, as the historical distributions of the macroeconomic factors are not necessarily normal, in order to improve the performance of LGD model each factor distribution is modelled by a Box-Cox transformation. We check the prediction accuracy of six base models with the different number of principal component of macroeconomic variables from 1 to 20. Between these 20 iterations, the best base model has the highest adjusted R-squared with eight principal components of macroeconomic factors. In model (1) we considered the bond characteristics as independent variables. In models (2) and (3) we added the first eight principal components of more than 100 macroeconomic variables but in model (2) we use the Box-Cox transformation.

Box and Cox (1964) present how to transform a variable to approximately normal distribution. They defined the following transformation:

$$\text{Transformed MacroFactor}=\begin{cases} \frac{(MacroFactor^{\lambda}-1)}{\lambda} & \lambda \neq 0 \\ log(MacroFactor) & \lambda = 0 \end{cases} \tag{3.15}$$

where λ is the unknown power transformation parameter. The fuzzy rule base is trained using the DE algorithm which is set to generate 15 rules. We ran the algorithm by setting different values to the number of rules from 5 to 40. Between all the 36 values in this range, the mean best results were obtained by the value of 15. Each rule refers to one partition in the input space and specifies a weight for each base model. Four types of defuzzification formulas defined in Section 3.3 –maximum formula, maximium of maximums formula, mean formula, and mean of maximum formula –are used to train the four types of fuzzy rules. The different results are compared. In order to apply the algorithm, all six base algorithms are trained and the final trained model is then stored.

Table 3.5 provides the mean and standard deviations of R^2, Adj. R^2, RMSE and

Table 3.4: Macroeconomic and financial predictor variables for principal components

Equity indexes and respective volatilities	
S&P 500	S&P 500 Vol
Russell 2000	Russell 2000 Vol
Nasdaq 100	Nasdaq 100 Vol
S&P small cap index	S&P small cap index Vol
Dow Jones industrial average Vol	
Cost of resource and capital	
PPI interm. energy goods	PPI crude energy materials
PPI industrial commodities	PPI all commodities
PPI finished goods	PPI intermediate materials
AAA corporate bond yield	BAA corporate bond yield
Corporate yield spread	Bank prime loan rate
30 year mortgage rate	S&P small cap index
Effective federal funds rate	Term structure spread
1 month commercial paper rate	3 month commercial paper rate
Wages/cost factors	
Employment cost index: wages & salaries	Employment cost index: benefits
Employee compensation index	Total wages & salaries
Management salaries	Durable manufacturing wages
Non-durable manufacturing wages	Unit labor cost: non-farm business
Unit labor cost: manufacturing	
Profitability measures	
Net corporate dividends	Corp. profits
Corporate net cash flow	After tax earnings
Business cycle and macro indicators	
New building permits	Housing starts
New houses sold	Non-farm housing starts
New orders: durable goods	New orders: capital goods
Inventory/sales-ratio	Capacity util. manufacturing
Change in private inventories	Capacity util. total industry
Inventories: total business	Light weight vehicle sales
Final sales to domestic buyers	Real GDP
Civilian employment	Employment/population-ratio
Unemployment rate	Unemployed, more than 15 weeks
Total no unemployed	Weekly hours worked
Private fixed investments	Real disposable personal income
ISM manufacturing index	Industr. production index
Uni Michigan consumer sentiment	Final sale of dom. product
National income	Personal Income
Manuf. industry output	Consumption expenditure
Manuf. industry production	Expenditure durable goods
Government expenditure	Gross private domestic investment
M2 money stock	CPI: all items less food
Personal savings	Personal savings rate
Gross saving	Uni Michigan¡ infl. expectations
CPI: energy index	GDP deflator, implicit
International Competitiveness	
Real exports goods, services	Real imports goods & services
Balance on merchandise trade	Trade weighted USD (Dollar Index)
FX index major trading partners	Current account balance
Financial Conditions	
Total loans and leases, all banks	Total commercial loans
Total consumer credit outst.	Total net loan charge-offs
Total borrowings from federal reserve	Total real estate loans
Household obligations/income	Commercial & industrial loans
Non-performing loans ratio	Non-performing loans ratio small banks
Non-perf. commercial loans	Net loan losses
Bank loans and investments	Household debt service payments
Federal debt of non-fin. industry	Loan loss reserves
Excess reserves of dep. institutions	Return on bank equity

MAE for the five folds cross-validation. The best performing model according to each metric is underlined in Table 3.5. Although all of the performance metrics listed above are useful measures, the most popular is the coefficient of determination R^2 or Adj. R^2 to compare model performance.

The Adj. R^2 of the basic model varies from about 2% to 29%. It can be observed that the maximum defuzzification fuzzy has the best accuracy performance in terms of R^2, Adj. R^2 and RMSE. This result is also consistent with the findings in Yao et al. (2015), Hartmann-Wendels et al. (2014), Qi and Zhao (2011), Loterman et al (2011) and Bastos (2010) who showed that non-parametric models outperformed the parametric models for LGD prediction.

Table 3.6 presents p-values of right side the Mann-Whitney U test for difference of Adj. R^2 and pair wise t-tests for RMSE and MAE differences between model (3) compared to model (1) and model (2). From the table it can be seen that there is significant outperformance for all advanced models compared to the 10 basic models for all performance metrics. Leow et al. (2014), Bellotti and Crook (2012), and Tobback et al (2014) report that the adding of macroeconomic factors improves the predictive performance of the LGD models.

Investigating the effects of the Box-Cox transformation of macroeconomic factors in LGD models, we find the performance of all fuzzy models become noticeably worse if the Box-Cox transformation is not made. However, Yao et al. (2015) report logistic and beta transformations of LGD do not improve prediction accuracy of SVR models. On the other hand, the prediction accuracy of the regression model was improved by using the Box-Cox transformation significantly.

In Tables 3.7 and 3.8 we provide the t-values of paired t-tests for differences between the RMSE and MAE between the different models with the independent variables of model (3) and model (2). Also, the p-values of the left side Mann-Whitney U test for difference of Adj. R^2 between the different model configurations are reported in Table 3.7 and 3.8. Fuzzy models have higher out-of-sample prediction accuracies compared to SVR models, regression trees and OLS regressions when applied to LGD modeling of US corporate bonds. In model (3), the mean for the maximum defuzzification approach in the fuzzy rule-based model illustrates higher predictive accuracy compared

Table 3.5: Performance measures from cross validation and the respective standard deviations for the models. The independent variables we have used in model (1) are basic variables of our model. In model (2) and (3), we add the first 8 principal components of 104 macroeconomic variables but in model (2), we use Box-Cox transformation. The best performing model according to each metric is underlined. (Reg. Tree: Regression Tree; Lin. Reg.: Linear Regression; $\epsilon - insensitive$: ϵ- Support Vector Regression; LS-SVR: Least Squared Support Vector Regression; LS-SVR DB: Least Squared Support Vector Regression with Different Intercepts; SP LS-SVR: Semi-Parametric Least Squared Support Vector Regression; Max DE Fuzzy: Maximum DE Fuzzy; Mean of Max DE Fuzzy: Mean of maximums DE Fuzzy; Max of Max DE Fuzzy: Maximum of maximums DE Fuzzy; Mean DE Fuzzy)

Model (1)	R^2	σ_{R^2}	*Adj.* R^2	$\sigma_{Adj.\,R^2}$	$RMSE$	σ_{RMSE}	MAE	σ_{MAE}
Reg. Tree	*0.2543*	*0.5136*	*0.2330*	*0.0305*	*25.8343*	*0.0296*	*21.3529*	*0.4832*
Lin. Reg.	*0.1725*	*0.2639*	*0.1490*	*0.0164*	*27.2167*	*0.0160*	*23.0692*	*0.2291*
$\epsilon - insensitive$	*0.0217*	*0.5318*	*0.0164*	*0.0365*	*29.7410*	*0.0354*	*24.3843*	*0.2440*
$LS - SVR$	*0.2267*	*0.3955*	*0.2046*	*0.0239*	*26.3105*	*0.0232*	*21.3177*	*0.3161*
$LS - SVR\,DB$	*0.2287*	*0.3910*	*0.2067*	*0.0236*	*26.2754*	*0.0229*	*21.2836*	*0.3139*
$SP\,LS - SVR$	0.2343	*0.4939*	*0.2125*	*0.0296*	*26.1780*	*0.0288*	*21.3155*	*0.3943*
$Max\,DE\,Fuzzy$	<u>0.2980</u>	0.4526	<u>0.2781</u>	0.0260	<u>25.0656</u>	0.0259	20.3001	0.4942
$Mean of Max\,DE\,Fuzzy$	0.2730	0.4927	0.2523	0.0288	25.5084	0.0280	<u>20.2374</u>	0.3703
$Max of Max\,DE\,Fuzzy$	0.2765	0.4861	0,2558	0.0284	25.4479	0.0277	20.9918	0.3088
$Mean\,DE\,Fuzzy$	0.2955	0.5277	0.2754	0.0305	25.1099	0.0296	20.2581	0.4507
Model (2)	R^2	σ_{R^2}	*Adj.* R^2	$\sigma_{Adj.\,R^2}$	$RMSE$	σ_{RMSE}	MAE	σ_{MAE}
Reg. Tree	*0.4113*	*0.6713*	*0.3791*	*0.0366*	*22.9507*	*0.0347*	*15.7887*	*0.5954*
Lin. Reg.	*0.4019*	*0.1279*	*0.3693*	*0.0070*	*23.1395*	*0.0066*	*18.4335*	*0.1211*
$\epsilon - insensitive$	*0.1875*	*0.6078*	*0.1432*	*0.0386*	*26.9657*	*0.0366*	*21.4519*	*0.2823*
$LS - SVR$	*0.5475*	*0.4048*	*0.5228*	*0.0192*	*20.1247*	*0.0182*	*13.8080*	*0.2025*
$LS - SVR\,DB$	*0.5539*	*0.4110*	*0.5296*	*0.0194*	*19.9814*	*0.0184*	*13.7305*	*0.2052*
$SP\,LS - SVR$	*0.6334*	*0.5277*	*0.6133*	*0.0227*	*18.1117*	*0.0215*	*12.4752*	*0.3047*
$Max\,DE\,Fuzzy$	0.6473	0.5044	0.6303	0.0210	17.7632	0.0200	12.1853	0.4147
$Mean of Max\,DE\,Fuzzy$	0.6425	0.5499	0.6252	0.0232	17.8843	0.0221	12.2273	0.2589
$Max of Max\,DE\,Fuzzy$	<u>0.6524</u>	0.3777	<u>0.6356</u>	0.0156	<u>17.6365</u>	0.0149	12.3118	0.4490
$Mean\,DE\,Fuzzy$	0.6425	0.3676	0.6251	0.0154	17.8883	0.0147	<u>12.1763</u>	0.4640
Model (3)	R^2	σ_{R^2}	*Adj.* R^2	$\sigma_{Adj.\,R^2}$	$RMSE$	σ_{RMSE}	MAE	σ_{MAE}
Reg. Tree	*0.4336*	*0.4915*	*0.4175*	*0.0254*	*22.5137*	*0.0247*	*15.1820*	*0.5507*
Lin. Reg.	*0.3734*	*0.2352*	*0.3555*	*0.0128*	*23.6840*	*0.0124*	*18.7640*	*0.1435*
$\epsilon - insensitive$	*0.1663*	*0.4577*	*0.1425*	*0.0287*	*27.3179*	*0.0279*	*21.7560*	*0.2093*
$LS - SVR$	*0.5172*	*0.3894*	*0.5035*	*0.0186*	*20.7865*	*0.0181*	*14.9632*	*0.2993*
$LS - SVR\,DB$	*0.5381*	*0.3867*	*0.5249*	*0.0181*	*20.3322*	*0.0176*	*14.4175*	*0.3162*
$SP\,LS - SVR$	*0.6492*	*0.6281*	*0.6392*	*0.0256*	*17.7132*	*0.0249*	*11.9941*	*0.3891*
$Max\,DE\,Fuzzy$	0.6602	0.5869	0.6438	0.0240	17.4332	0.0229	11.8842	0.5871
$Mean of Max\,DE\,Fuzzy$	<u>0.6621</u>	0.6892	<u>0.6457</u>	0.0283	<u>17.3832</u>	0.0270	<u>11.7114</u>	0.4764
$Max of Max\,DE\,Fuzzy$	0.6617	0.5088	0.6453	0.0208	17.3983	0.0198	12.0072	0.4217
$Mean\,DE\,Fuzzy$	0.6597	0.4430	0.6432	0.0182	17.4495	0.0173	11.7981	0.5213

Table 3.6: Paired t-test for differences of the RMSE and MAE between the respective models comparing the models using independent variables model (3) in Table 4 with the models using independent variables from models (1) and (2) in Table 4. The p-values of the right side Mann-Whitney U test for difference of Adj. R^2 between the same models report. A small p-value indicates that the model (3) Adj. R^2 of related technique is higher than the model (1) or (2) and a p-value nearby 1 indicates that the model (3) Adj. R^2 of related technique is significantly lower than the model (1) or (2). Statistical significance on the 99% level is indicated with ** and significance on the 95% level is indicated with *.

Model (1)	$Reg.\,Tree$	$Lin.\,Reg.$	$\epsilon - insensitive$	$LS-SVR$	$LS-SVR\,DB$
RMSE	-9.7969**	-101.1875**	-13.4966**	-38.2998**	-42.086**
MAE	-18.6766**	-99.2179**	-21.5918**	-49.6014**	-47.6485**
Adj. R^2	0.00397**	0.00397**	0.00397**	0.00397**	0.00397**
Model (1)	$SP\,LS-SVR$	$Max\,DE\,Fuzzy$	$MeanOfMax\,DE\,Fuzzy$	$MaxOfMax\,DE\,Fuzzy$	$Mean\,DE\,Fuzzy$
RMSE	-46.0092**	-31.9176**	-25.6815**	-35.2149**	-32.4535**
MAE	-46.592**	-22.9743**	-30.7784**	-46.0412**	-28.493**
Adj. R^2	0.00397**	0.00397**	0.00397**	0.00397**	0.00397**
Model (2)	$Reg.\,Tree$	$Lin.\,Reg.$	$\epsilon - insensitive$	$LS-SVR$	$LS-SVR\,DB$
RMSE	-1.0237	9.4929**	4.154*	6.5802**	3.8881*
MAE	-1.2916	8.6686**	7.7986**	15.0589**	10.8452
Adj. R^2	0.2103	1(**)	0.8889	0.9921(**)	0.8452
Model (2)	$SP\,LS-SVR$	$Max\,DE\,Fuzzy$	$MeanOfMax\,DE\,Fuzzy$	$MaxOfMax\,DE\,Fuzzy$	$Mean\,DE\,Fuzzy$
RMSE	-2.5936	-2.3624	-2.022	-2.079	-5.1372**
MAE	-5.2204**	-2.3962	-2.659	-2.0798	-3.512*
Adj. R^2	0.2738	0.2103	0.1111	0.1548	0.1111

to the other SVR models.

Table 3.9 presents the five variables with the biggest influence on the first eight principal components of the 104 macroeconomic variables. The first principal component mainly incorporates housing indicators. As our sample period includes the financial and housing crisis of 2007/8 we can interpret the first principal component as a crisis indicator. The second principal component mainly incorporates information about the real economy from an output perspective while the third principal component includes mainly information about corporate profits. The fourth component is driven mainly on interest rates and inflation variables although a mix of variables contributes to the fifth component. The sixth principal component includes mostly by the stock markets and the other principal components are a mix of macroeconomic variables. However, it is difficult to interpret the large number of economic indicators that are entering the model.

As we do not have access to the database of other researchers, it is not possible to compare predictive abilities of our results with their results. We show that the fuzzy decision fusion approach has higher prediction accuracy than SVR models used by Yao et al. (2015). For example, best performance of our proposed model on the dataset with principal components of macroeconomic variables is for Mean of Max DE fuzzy with R-squared 0.662. The corresponding best performance for single models is for SP LS-SVR with R-squared 0.649. The closest LGD distribution from the study Loterman et al. (2012) to our data set is their fifth data set. They reported non-parametric methods outperform for LGD analysis and the highest R-squared value is 0.3486. Loterman et al. (2012) reported that the best model for the distribution of a certain shape may not necessarily be the best for other distributions. As our LGD distribution is different from the study by Yao et al. (2015), we cannot say that fuzzy fusion techniques would have better prediction accuracy than SVRs for their data. However, this improvement also comes with some practical limitations. As mentioned before, all base models are needed to be trained and stored before training the fuzzy model. The number of rules in the rule base is also one of the parameters which needs to be tuned in advanced. This limitation could be overcome, not omitted by using some more advanced algorithms which tune the number of rules along with training the fuzzy rule base.

Table 3.7: Paired t-tests for differences of the RMSE and MAE between the different models with the independent variables of model (3) in Table 3.5. For RMSE and MAE the value of the t-statistic is positive when the model in the row is better than the model in the respective column. The p-values of the left side Mann-Whitney U test for difference of Adj. R^2 between the same models are reported. A small p-value indicates that the Adj. R^2 of the model of the related row is higher than the Adj. R^2 of the model of related column and vice versa. Statistical significance on the 99% level is indicated with ** and significance on the 95% level is indicated with *.

RMSE	1	2	3	4	5	6	7	8	9
Reg. Tree (1)									
Lin. Reg. (2)	-0.62								
$\epsilon - insensitive$ (3)	-12.97**	-16.13**							
$LS - SVR$ (4)	6.80**	22.52**	28.25**						
$LS - SVR\ DB$ (5)	7.07**	23.06**	27.99**	16.52**					
$SP\ LS - SVR$ (6)	10.63**	23.83**	26.41**	12.41**	11.90**				
$Max\ DE\ Fuzzy$ (7)	11.82**	28.15**	30.92**	18.89**	18.44**	7.45**			
$Mean of Max\ DE\ Fuzzy$ (8)	11.36**	23.34**	27.44**	11.80**	11.25**	4.96**	-1.63		
$Max of Max\ DE\ Fuzzy$ (9)	13.45**	40.98**	31.73**	22.74**	22.45**	5.18**	1.68	2.04	
$Mean\ DE\ Fuzzy$ (10)	13.06**	40.95**	32.30**	21.54**	20.99**	2.14	-1.74	-0.03	-17.81**
MAE	1	2	3	4	5	6	7	8	9
Reg. Tree (1)									
Lin. Reg. (2)	-10.09**								
$\epsilon - insensitive$ (3)	-16.90**	-18.95**							
$LS - SVR$ (4)	5.67**	42.80**	56.16**						
$LS - SVR\ DB$ (5)	5.89**	43.11**	54.64**	11.90**					
$SP\ LS - SVR$ (6)	8.64**	42.26**	46.83**	23.22**	23.61**				
$Max\ DE\ Fuzzy$ (7)	9.62**	34.49**	35.82**	11.83**	11.75**	2.86*			
$Mean of Max\ DE\ Fuzzy$ (8)	10.60**	76.31**	47.92**	16.06**	15.18**	1.54	-0.27		
$Max of Max\ DE\ Fuzzy$ (9)	8.32**	33.57**	34.43**	10.83**	10.52**	1.82	-1.31	-0.73	
$Mean\ DE\ Fuzzy$ (10)	8.51**	31.28**	36.63**	12.10**	11.84**	3.27*	0.12	0.34	2.11
Adj. R^2	1	2	3	4	5	6	7	8	9
Reg. Tree (1)									
Lin. Reg. (2)	1(**)								
$\epsilon - insensitive$ (3)	1(**)	1(**)							
$LS - SVR$ (4)	0.004**	0.004**	0.004**						
$LS - SVR\ DB$ (5)	0.004**	0.004**	0.004**	0.075					
$SP\ LS - SVR$ (6)	0.004**	0.004**	0.004**	0.004**	0.004**				
$Max\ DE\ Fuzzy$ (7)	0.004**	0.004**	0.004**	0.004**	0.004**	0.274			
$Mean of Max\ DE\ Fuzzy$ (8)	0.004**	0.004**	0.004**	0.004**	0.004**	0.210	0.5		
$Max of Max\ DE\ Fuzzy$ (9)	0.004**	0.004**	0.004**	0.004**	0.004**	0.210	0.421	0.727	
$Mean\ DE\ Fuzzy$ (10)	0.004**	0.004**	0.004**	0.004**	0.004**	0.273	0.5	0.790	0.579

Table 3.8: Paired t-tests for differences of the RMSE and MAE between the different models with the independent variables of model (2) in Table 3.5. For RMSE and MAE the value of the t-statistic is positive when the model in the row is better than the model in the respective column. The p-values of the left side Mann-Whitney U test for difference of Adj. R^2 between the same models are reported. A small p-value indicates that the Adj. R^2 of the model of the related row is higher than the Adj. R^2 of the model of related column and vice versa. Statistical significance on the 99% level is indicated with ** and significance on the 95% level is indicated with *.

RMSE	1	2	3	4	5	6	7	8	9
Reg. Tree (1)									
Lin. Reg. (2)	-5.41**								
$\epsilon - insensitive$ (3)	-15.07**	-21.13**							
$LS - SVR$ (4)	6.25**	23.61**	99.73**						
$LS - SVR\,DB$ (5)	8.21**	29.19**	82.93**	23.58**					
$SP\,LS - SVR$ (6)	15.01**	24.89**	34.32**	13.71**	12.53**				
$Max\,DE\,Fuzzy$ (7)	17.42**	28.71**	38.94**	17.00**	15.99**	7.46**			
$Mean of Max\,DE\,Fuzzy$ (8)	15.42**	23.28**	28.43**	11.77**	10.82**	3.81*	0.44		
$Max of Max\,DE\,Fuzzy$ (9)	18.83**	36.69**	45.63**	21.63**	21.03**	4.12*	0.69	-0.11	
$Mean\,DE\,Fuzzy$ (10)	21.46**	44.41**	43.44**	20.54**	19.97**	2.46	-0.19	-0.47	-1.02
MAE	1	2	3	4	5	6	7	8	9
Reg. Tree (1)									
Lin. Reg. (2)	-12.93**								
$\epsilon - insensitive$ (3)	-30.57**	-26.14**							
$LS - SVR$ (4)	1.137	31.94**	60.06**						
$LS - SVR\,DB$ (5)	2.16	32.77**	58.36**	18.49**					
$SP\,LS - SVR$ (6)	28.30**	36.30**	59.18**	30.39**	26.35**				
$Max\,DE\,Fuzzy$ (7)	13.90**	27.43**	34.71**	15.49**	13.56**	0.68			
$Mean of Max\,DE\,Fuzzy$ (8)	20.81**	32.93**	43.91**	23.01**	20.78**	3.60*	1.64		
$Max of Max\,DE\,Fuzzy$ (9)	25.32**	34.49**	52.64**	27.88**	24.60**	-0.53	-0.86	-5.25**	
$Mean\,DE\,Fuzzy$ (10)	23.29**	29.89**	44.11**	23.81**	22.05**	2.18	0.59	-1.39	4.02*
Adj. R^2	1	2	3	4	5	6	7	8	9
Reg. Tree (1)									
Lin. Reg. (2)	0.789								
$\epsilon - insensitive$ (3)	1(**)	1(**)							
$LS - SVR$ (4)	0.004**	0.004**	0.004**						
$LS - SVR\,DB$ (5)	0.004**	0.004**	0.004**	0.274					
$SP\,LS - SVR$ (6)	0.004**	0.004**	0.004**	0.004**	0.004**				
$Max\,DE\,Fuzzy$ (7)	0.004**	0.004**	0.004**	0.004**	0.004**	0.210			
$Mean of Max\,DE\,Fuzzy$ (8)	0.004**	0.004**	0.004**	0.004**	0.004**	0.210	0.655		
$Max of Max\,DE\,Fuzzy$ (9)	0.004**	0.004**	0.004**	0.004**	0.004**	0.111	0.421	0.274	
$Mean\,DE\,Fuzzy$ (10)	0.004**	0.004**	0.004**	0.004**	0.004**	0.274	0.726	0.5	0.845

Table 3.9: The five variables with the biggest influence on each of the 8 principal components

Variable	PC1	Variable	PC2
Nonperforming Loans to Total Loans	0.1398	Real Gross Domestic Product, 3 Decimal	-0.1803
Housing Starts: Total: New Privately Owned Housing Units Started	-0.1396	Real Final Sales of Domestic Product	-0.1653
New Private Housing Units Authorized by Building Permits	-0.1378	Producer Price Index by Commodity Intermediate Energy	-0.1650
Housing Starts: Total: New Privately Owned Housing Units Started	-0.1367	Real imports of goods and services	-0.1648
New One Family Houses Sold: United States	-0.1357	Producer Price Index by Commodity Industrial Commodities	-0.1638
Variable	PC3	Variable	PC4
ISM Manufacturing: PMI Composite Index	-0.2290	Moody's Seasoned Aaa Corporate Bond Yield	0.2589
Corporate Profits After Tax (with IVA and CCAdj)	-0.2087	University of Michigan Inflation Expectation	0.2358
Corporate Profits After Tax (without IVA and CCAdj)	-0.2027	10-Year Treasury Constant Maturity Rate	0.2277
Household Financial Obligations as a percent of Disposable Personal Income	0.1993	Gross Saving	-0.2032
Commercial and Industrial Loans, All Commercial Banks	0.1973	30-Year Conventional Mortgage Rate	0.1974
Variable	PC5	Variable	PC6
NASDAQ 100 Index	0.3414	Moody's Seasoned Baa Corporate Bond Yield	0.5792
Moody's Seasoned Baa Corporate Bond Yield	-0.3097	NASDAQ 100 Index	-0.5234
TermStructure	-0.2053	Vol S&P 500	-0.2126
University of Michigan Inflation Expectation	-0.2043	Russell 2000 Vol 1m	-0.1815
Producer Price Index by Commodity for Crude Energy Materials	-0.1787	CBOE NASDAQ 100 Volatility Index	-0.1780
Variable	PC7	Variable	PC8
University of Michigan Inflation Expectation	0.3203	Personal Saving Rate	0.2726
Total Borrowings of Depository Institutions from the Federal Reserve	-0.2708	Corporate Profits After Tax (without IVA and CCAdj)	-0.2567
Personal Saving	0.2236	Corporate Profits After Tax (with IVA and CCAdj)	-0.2544
Personal Saving Rate	0.2184	Russell 2000 Vol 1m	-0.2515
Nonperforming Commercial Loans to Commercial Loans	-0.1840	Vol S&P 500	-0.2413

The performance of all techniques improves by adding principal components of macroeconomic variables to the LGD model as explanatory variables in model (2) and model (3). For example, the Adj. R^2 of regression model increases from 15% to 36% just by adding the principal components of macroeconomic factors.

3.6 Conclusions

To the best of our knowledge, there is no study that applies and compares fuzzy techniques for predicting the LGD of corporate bonds. In this paper, we compare the predictive performance of fuzzy techniques with SVR methods, regression trees and OLS regressions to predict corporate bond LGD. Fuzzy rule-based models have shown to be strong function approximators. Our findings suggest that fuzzy rule-based models are more accurate than other methods identified in the literature for predicting LGD for defaulted corporate bonds. Adding the principal components derived from 104 macroeconomic measures improve the predictive accuracy of the SVR and fuzzy models. We use a meta heuristic DE algorithm to create an optimized fuzzy rule base with an appropriate number of rules to deal with the complex benchmark data. Moreover, although the Box-Cox transformation of macroeconomic factors improves the accuracy of the regression model, it does not improve the performance accuracy of fuzzy techniques.

The results reported in this paper suggest more accurate ways for computing the regulatory capital required by Basel Accords for banks searching for a more precise method to predict LGD for corporate bonds.

Chapter 4

Intertemporal defaulted bond recoveries prediction via machine learning

This chapter is joint work with Dr. Abdolreza Nazemi[1] and Prof. Frank J. Fabozzi[2] from the unpublished working paper with the same title.[3]

4.1 Introduction

Under the advanced internal ratings-based approach the Basel II/III accords allow financial institutions to use their own estimates for the credit risk parameters. Thus, accurate and reliable estimates of the credit risk parameters probability of default, recovery rate, and exposure at default are needed. Traditionally, in credit risk analysis much attention has been paid to the probability of default while the recovery rate has been set to constant values not taking into account its time variation and its cross-sectional variation. In particular, the time variation of recovery rates has been neglected in the literature.

The most recent studies have examined out-of-sample or in-sample settings to analyze the determinants of recovery rates. According to Kalotay and Altman (2017) the applicability of out-of-sample estimation to the field of recovery rate prediction is questionable. In particular, k-fold cross validation is commonly used for performance

[1] School of Economics and Business Engineering, Karlsruhe Institute of Technology
[2] EDHEC Business School, Nice, France
[3] Section 4.6 is not part of this working paper.

measurement. During k-fold cross validation the dataset is randomly divided into k subsamples. Each subsample is used for out-of-sample prediction once while the respectively remaining k-1 subsamples are used for training. The performance measurement is obtained as average of the predictions for the k-th subsample.

Even though both out-of-time and out-of-sample estimation make a distinction between training and testing data, out-of-sample estimation for recovery rate prediction suffers from two shortcomings. First, as the dataset is randomly divided into partitions during out-of-sample estimation it is likely that there observations used for training the model that have occured after the observations used for testing the model. So, the data-generating process is assumed to be time invariant. Second, it is questionable whether the recovery rates of two defaulted bonds issued by the same company are independent of each other. Considering the case when two bonds from the same issuer have defaulted at the same time, only during out-of-time estimation these two bonds are together either both in the training set or both in the test set. We address these challenges by comparing a wide range of statistics and machine learning methods such as inverse Gaussian regression, random forest, sparse power expectation propagation, and support vector regression not only for out-of-sample but also for out-of-time prediction of recovery rates on defaulted corporate bonds.

There has been increasing application of machine learning techniques to finance since the turn of the century.[4] The volume, variety, velocity, and veracity of available financial data have increased for several reasons such as improvements in computational and storage power. Moreover, the rise of the Internet has made available large sets of data that allow researchers to use and merge them for different purposes and to automatically store the associated data for many financial transactions.

This study includes a large number of macroeconomic variables relating to corporate bond recovery recovery rates. We compare selection techniques such as the stability selection, the MC+ algorithm, and the SparseStep algorithm for selecting the subset of the macroeconomic variables which is most related to the recovery rate. This paper is the first study that compares these econometrics and machine learning methods in empirical finance. Furthermore, we extend our analysis to include alternative data sources

[4] See, for example, Fuster et al. (2017), Kleinberg et al. (2017), Jean et al. (2016), Manela et al. (2017), Gu at al. (2018), and Giannone et al. (2017).

such as text-based measures from front-page articles of the *Wall Street Journal* as independent variables. By including text-based measures of investors' uncertainty we add further macroeconomic and uncertainty information. Our study is the first paper to use uncertainty measures from news for the prediction of corporate bond recovery rates.

In this study, we contribute to the recovery rate literature in four ways. First, in addition to present a machine learning framework for out-of-sample recovery rate prediction, this study evaluates the intertemporal prediction performance of a wide range of parametric and non-parametric techniques that surprisingly have attracted less attention than out-of-sample prediction in the literature. Our findings demonstrate that machine learning techniques deliver superior predictive performance compared to traditional techniques not only out-of-sample but also out-of-time.

Second, we include news-based variables as an alternative group of independent variables in our analysis. By incorporating these text-based variables, we show that news-based variables are significant drivers for recovery rate estimation. Third, we employ high-dimensional data and select the most informative macroeconomic variables using several selection techniques. By comparing the out-of-sample performances of these techniques, we find that the new machine learning method (SparseStep) outperform other methods. Lastly, we investigate the importance of the groups of variables by ranking all independent groups of variables with a random forest.

We organize the remainder of the paper as follows. A review of the literature is presented in Section 4.2. In Section 4.3 we describe the models and selection algorithms we applied. We describe the data we used in Section 4.4 and we present our empirical results in Section 4.5. We investigate the behaviour of recovery rates under macroeconomic stress in Section 4.6. Our paper is concluded in Section 4.7.

4.2 Literature review

Krüger and Rösch (2017) study the downturn loss-given-default employing the quantile regression technique for both in-sample and out-of-sample estimation. Nazemi and Fabozzi (2018) report that support vector regression techniques outperform other methods for predicting recovery rates of U.S. corporate bonds in an out-of-sample study.

Altman and Kishore (1996) show that the defaulted debt from public utilities (70%) and chemical, petroleum, and related products (63%) exhibits the highest average recovery rates. Moreover, they find that controlling for the seniority the original rating of a defaulted bond has no impact on the recovery rate. Acharya et al. (2007) document that creditors recover less if the industry of the defaulted firm is in distress. In particular, they show on a dataset from 1982 to 1999 that defaulted corporate bonds in distressed industries exhibit 10% to 15% lower average recovery rates. Altman et al. (2005) find that default rates, seniority, and collateral levels are important determinants of recovery rates of corporate bonds. Focusing on the macroeconomic determinants of recovery rates they find that while there is a significant negative relationship between realized default rates and recovery rates, other macroeconomic variables such as the growth rate of the gross domestic product and the return of the stock market have only weak correlation with the average recovery rate.

Altman and Kalotay (2014) introduce a modeling approach based on mixtures of Gaussian distributions conditioned on borrower characteristics, instrument characteristics, and credit market conditions. They show that the forecasts generated by this method are more accurate than parametric regression-based forecasts during out-of-time estimation. Jankowitsch et al. (2014) examine the recovery rates of defaulted bonds while paying special attention to the trading microstructure around various types of default events in an in-sample study. They find that (1) high trading volumes on the default day and the following 30 days with reduced trading activity after this time period and (2) bond characteristics (e.g. coupon, CDS availability, and covenants) have a significant impact on market-based recovery rates in an in-sample analysis.

Qi and Zhao (2011) find that non-parametric techniques such as regressions trees and neural networks outperform parametric methods both in-sample and out-of-sample for the prediction of corporate bonds' recovery rates. Yao et al. (2015) argue that accounting for the heterogeneity of bond seniorities within least squares support vector techniques enhances their predictive capacity for recovery rates of corporate bonds in an out-of-sample estimation.

Bastos (2014) illustrates how ensembles of models derived from the same regression method yield more accurate forecasts of recovery rates than a single model. In partic-

ular, using bootstrap aggregation (bagging) to build an ensemble of regression trees, he shows that his results are valid for both corporate bonds and loans both during out-of-sample estimation and cross validation.

Chen (2010) states that the average values of recovery rates during the recessions (1982, 1990, 2001, and 2008) were smaller than during economic upswings. Bruche and Gonzales-Aguado (2010) argue that in recessions more firms default while the average recovery rate decreases. They propose an econometric model incorporating the credit cycle as unobserved Markov chain to account for time variation in the probability of default and the recovery rate. They conclude that the time-variation in recovery rate distributions amplifies risk.

Calabrese and Zenga (2010) suggest a beta regression model for the estimation of bank loans recovery rates. Hartmann-Wendels et al. (2014) forecast recovery rates on a dataset of defaulted leasing contracts provided by three German leasing companies. In their study model trees outperform regression-based approaches out-of-sample. They emphasize the importance of out-of-sample estimation for appropriate risk management. Cheng and Cirillo (2018) investigate a nonparametric survival approach to estimate the recovery rate and recovery time of private loans.

Mora (2015) argues that macroeconomic conditions do matter for recovery rate prediction. She shows how recovery rates in different industries are impacted by macroeconomic conditions in different ways. Studies such as Cantor and Varma (2004), Acharya et al. (2007), Qi and Zhao (2011), Jankowitsch et al. (2014), and Yao et al. (2015) use only a few macroeconomic variables. Nazemi and Fabozzi (2018) investigate the relationship between recovery rates of corporate bonds and macroeconomic variables out-of-sample. They implemented the least absolute shrinkage and selection operator (LASSO) for determining the most relevant macroeconomic variables from a comprehensive macroeconomic data set to recovery rates. The models including the macroeconomic variables selected by LASSO outperform the models including a few macroeconomic variables which are typically used in the literature on recovery rates.

Compelled by the sparse literature on out-of-time estimation of recovery rates Kalotay and Altman (2017) investigate the time variation of recovery rates. Comparing cross-sectional and intertemporal predictive performance they conclude that

machine learning techniques such as the regression tree fail to outperform traditional techniques such as inverse Gaussian regression in an intertemporal setting. Further, applying conditional mixture models they improve estimates of expected credit losses by taking the time variation of the recovery rate distribution into account. A fast maximum-likelihood approach for the estimation of conditional mixtures of distributions enables their analysis.

Our study is similar to the recent work of Kalotay and Altman (2017) published in this journal. In contrast to their research, we find that machine learning techniques outperform also during intertemporal prediction. Furthermore, we include text-based variables in the analysis, select the most informative predictors from 182 macroeconomic variables, and investigate the permutation importance of the groups of explanatory variables.

4.3 Corporate bond recovery rate modeling

In this section, a description of the modeling techniques this study uses for recovery rate prediction and for the selection of macroeconomic variables is provided. We use the built-in implementations in MATLAB for the models presented in 4.3.1-4.3.4.

4.3.1 Linear regression as benchmark

For comparing the out-of-time and out-of-sample performance of our machine learning techniques with more statistical methods, we include a traditional linear regression model as a benchmark model. Thus, we estimate the following linear regression model for the recovery rate r_{ijn} of bond i in industry j at the time of default of the n-th bond to serve as a benchmark for our machine learning models:

$$\begin{aligned} r_{ijn} =& \alpha + \beta_c(\text{instrument-specific variables})_i \\ &+\nu(\text{industry distress variables})_{jn} \\ &+\eta(\text{news-based variables})_n \\ &+\zeta(\text{macroeconomic variables})_n \\ &+\epsilon_{ijn} \quad \epsilon_{ijn} \sim N(0, \sigma^2) \end{aligned} \tag{4.1}$$

We control for the instrument-specific variables with dummy variables for the industry, the seniority, the coupon type, and the instrument type. The industry distress variables indicate whether the performance of the industry index was worse than -30% and the sales growth was negative in the year preceding the default. The news-implied volatility variables are text-based measures capturing uncertainty and disaster risk. The various methodologies for selecting the macroeconomic variables will be presented in section 4.2.7.

4.3.2 Inverse Gaussian regression

Due to its popularity in recovery rate modeling in studies such as Qi and Zhao (2011) and Kalotay and Altman (2017), we also consider inverse Gaussian regression. In doing so, the recovery rates are transformed from the interval (0,1) to (-∞, ∞) using the inverse Gaussian cumulative distribution function. These transformed recovery rates are then regressed on the independent variables as described for the case of ordinary linear regression. Finally, the estimated values are transformed back from (-∞, ∞) to (0,1) using the Gaussian distribution function.

4.3.3 Regression tree

One class of machine learning methods that has been found to deliver very good predictive performance as well as an easy-to-understand model is the regression tree. Qi and Zhao (2011), Kalotay and Altman (2017), and Nazemi and Fabozzi (2018) have used regression trees successfully for LGD modeling. Two other advantages of the regression tree are that it can be used to model non-linearity and it exhibits a relatively robust behavior against outliers. For these reasons, we apply the classification and regression technique (CART) algorithm as defined by Breiman et al. (1984) for the creation of the regression tree model.

4.3.4 Random forest

Breiman (2001) introduces random forest as a model that is more robust and has a better predictive capacity out of sample than the regression tree. Random forest is

an improvement of bagging, which trains a large number of regression trees and then predicts the average of the trees' predictions. Better performance and reduced variance of the predictions are the advantages of bagging compared with regression trees. In a random forest, a random subset of explanatory variables is selected for each regression tree. The random forest has three tuning parameters: The minimum leaf size of the trees, the number of trees, and the number of explanatory variables used for each tree. We use one third of all explanatory variables for each tree in accordance with the default value from Breiman (2001). The number of trees and the minimum leaf size are determined by ten-fold cross validation on the training set.

4.3.5 Semiparametric least-squares support vector regression

Suykens and Vandewalle (1999) introduced a least-squares version of the support vector machine classifier. Enticed by the promising results from a study by Nazemi and Fabozzi (2018), we make use of a semiparametric least-squares support vector regression (SP LS-SVR) model which assumes the impact from the S different seniority classes to be linear.[5] The parameter C regularizes the quadratic errors u_{sj}^2 while N denotes the number of defaulted bonds and $\mathbf{W}$ denotes the weight vector of the independent variables. The kernel function for the feature mapping into the higher dimensional space is defined as $\phi(X_i)$ while the kernel matrix $\mathbf{K}$ is defined as $K(X_i, X_j) = \phi(X_i) \cdot \phi(X_j)$. β is a vector of fixed effects for the seniority of the respective group and the dummy variables for the seniority classes are denoted by z_{sj}.

$$\begin{aligned} \min\ & J(W, b, u_i) = \frac{1}{2}||W||^2 + \frac{1}{2}\beta^T\beta + \frac{1}{2}b^2 + \frac{C}{2}\sum_{s=1}^{S}\sum_{j=1}^{n_s} u_{sj}^2 \\ \text{s.t. } & r_i = \mathbf{W}^T\phi(X_i) + b + \beta^T z_{sj} + u_{sj}, \qquad j = 1, ..., n_s, s = 1, ..., S \end{aligned} \tag{4.2}$$

The Lagrangian function of this optimization problem evaluates to

$$L(W, b, u_{sj}, \alpha_{sj}) = J(W, b, u_{sj}) - \sum_{s=1}^{S}\sum_{j=1}^{n_s} \alpha_{sj}(W^T\phi(X_{sj}) + b + \beta^T z_{sj} + u_{sj} - r_{sj}) \tag{4.3}$$

Therefore, with $\mathbf{V}$ denoting a $N \times N$-matrix of ones and $\mathbf{Z}_{ij} = z_{sj}^T z_{sj}$, the dual formulation is

[5] We implement SP LS-SVR in MATLAB.

$$\min \frac{1}{2}\alpha^T \mathbf{K}\alpha + \frac{1}{2}\alpha^T \mathbf{Z}\alpha + \frac{1}{2}\alpha^T \mathbf{V}\alpha + \frac{1}{2C}\alpha^T \alpha - r^T \alpha \tag{4.4}$$

4.3.6 Power expectation propagation

According to Bui et al. (2017), Gaussian processes are flexible distributions over functions that are used for a wide range of applications such as regression, representation learning, and state space modeling. Bui et al. (2017) introduce a unifying framework for sparse Gaussian process pseudo-point approximation using power expectation propagation.[6] Their novel approach to sparse Gaussian process regression, a power expectation propagation framework, subsumes expectation propagation and the sparse variational free energy method into a unified framework for pseudo-point approximation.

In particular, if power expectation converges, its updates are equivalent to the original expectation propagation procedure while substituting the Kullback-Leibler divergence minimization with an alpha-divergence minimization. As alpha $\rightarrow 0$, the power expectation propagation solution becomes the minimum of a variational free energy approach. In contrast, when alpha $= 1$, the solution from the original expectation propagation approach is recovered. Bui et al. (2017) show that their innovative algorithm for Gaussian process regression outperforms both expectation propagation and variational free energy approaches. To the best of our knowledge, our paper is the first to apply sparse power expectation propagation in credit risk.

4.3.7 Selection of macroeconomic variables

In this study's analysis, we include 182 macroeconomic variables to account for time variation in the recovery rates due to macroeconomic changes. A broad range of variables from categories such as international competitiveness, as well as stock market, credit market, micro-level, and business cycle conditions are taken into consideration. We compare three methods to select the most informative macroeconomic variables: The stability selection, the SparseStep algorithm, and the MC+ algorithm.

[6] We use the MATLAB implementation from Bui et al. (2017) of the algorithm for power expectation propagation.

4.3.7.1 Least absolute shrinkage and selection operator

Tibshirani (1996) introduces LASSO, a regularized least squares method imposing a penalty on the L_1 norm of the regression coefficients. The LASSO estimates the regularized coefficients $\hat{B}$ as follows:

$$\{\hat{B}\} = \arg\min_{B} \|r - XB\|_2^2 + \lambda\|B\|_1 \tag{4.5}$$

where λ denotes the non-negative LASSO regularization coefficient and B denotes the LASSO regularization loadings and X_n denotes the $N \times 1$ vector $X_n = (x_1, ..., x_N)'$ of macroeconomic variables. By selecting a subset of the macroeconomic variables and eliminating the rest of the variables, the resulting model becomes more interpretable and exhibits a higher out-of-sample predictive accuracy than the complete model. In particular, Nazemi and Fabozzi (2018) have noted that recovery models with macroeconomic variables selected by LASSO outperform models with few macroeconomic variables.

As shown by Meinshausen and Bühlmann (2010), the variable selection by a LASSO regression can change with a small perturbation of the data. To address this issue, they introduce stability selection, which entails subsampling and evaluating the selection probability of each variable. Using stability selection, the variable selection is conducted repeatedly on random samples of the dataset and the number of times each variable is selected during this process is counted. Only the variables that have been selected with a higher relative frequency than the specified counting proportion are ultimately selected.[7] The main goal of this approach is to model high-dimensional data through a stable selection of the macroeconomic variables that capture the most information for recovery rate estimation. We check the robustness of the LASSO variable selection by applying stability selection with a counting proportion equal to 0.6.

4.3.7.2 SparseStep

Van den Burg et al. (2017) present the SparseStep algorithm. While LASSO penalizes the L_1 norm the SparseStep algorithm imposes a penalty on the counting norm L_0. Van den Burg et al. (2017) apply the following approximation to the counting norm

[7] We make use of the scikit-learn package in *Python* for the stability selection algorithm.

L_0:

$$\|\beta_l\|^0 \approx \frac{\beta_l^2}{\beta_l^2 + \gamma^2} \tag{4.6}$$

where γ denotes a positive constant, β_l denotes the l-th coefficient, and p is the number of independent variables. To arrive at a sparse solution the approximation to the exact counting norm L_0 is added for regularization:

$$\{\hat{\beta}\} = \arg\min_{\beta} \|r - X\beta\|_2^2 + \lambda \sum_{l=1}^{p} \frac{\beta_l^2}{\beta_l^2 + \gamma^2} \tag{4.7}$$

While LASSO is a biased estimator the SparseStep algorithm yields unbiased estimates of the parameter vector. Further, Van den Burg et al. (2017) argue that SparseStep often outperforms earlier approaches such as ridge regression or LASSO in both model fit and prediction accuracy.[8]

4.3.7.3 MC+ algorithm

Zhang (2010) introduces MC+ for penalized variable selection in high-dimensional linear regressions.[9] This method is based on two elements: a minimax concave penalty and a penalized linear unbiased selection algorithm. While the LASSO estimates are biased, MC+ provides nearly unbiased estimates. Zhang (2010) outlines the theoretical and empirical advantages of MC+ compared to LASSO; in particular, the increased selection accuracy of MC+ in a simulation setting.

4.3.8 Ranking variables by permutation importance

Altmann et al. (2010) outline how the conventional feature importance from random forests based on the mean decrease of impurity is biased towards categorical predictors with a large number of categories. In particular, they show that permutation importance is an importance measure that does not suffer from this bias.[10] Permutation importance is based on the mean decrease in accuracy and is computed as the difference between the baseline R^2 of the model and the R^2 of the model when one variable's (or group of variables') values are permuted randomly. Strobl et al. (2008) show that

[8] For the SparseStep algorithm we make use of package 'sparsestep' in *R*.
[9] We use the package 'plus' in *R* for the MC+ algorithm.
[10] We use the implementation of permutation importance from the Python package 'pimp'.

permutation importance suffers from a bias towards correlated variables. Building groups of variables instead of investigating the importance of each variable on its own enables us to generate a ranking that will suffer less from the multicollinearity inherent to our high-dimensional data. Following Gregorutti et al. (2015) we adjust the importance of each group by dividing it by the number of variables in the respective group.

4.4 Data

As illustrated in Figure 4.1, we merge several data sources such as S&P Capital IQ, Bloomberg, Federal Reserve Bank of St. Louis, and news from front-page article of the Wall Street Journal to analyze the recovery rate of U.S. corporate bonds in this study. Our initial data set consists of 2080 bonds that have defaulted between 2001 and 2016 retrieved from the S&P Capital IQ database (Capital IQ). The bond data are retrieved from S&P Capital IQ. All bonds are denominated in US dollar. Industry variables are retrieved from Bloomberg (BBG). A default event occurs when a company files for a Chapter 11 bankruptcy petition or is assigned a rating of 'D' (meaning that the debtor is in default) or 'SD' (selective default) by Standard Poor's. The issuers of our bonds can be assigned to the following industries: industry, consumer discretionary, consumer staples, telecommunications, raw materials, utilities, energy, financial services and information technology.

Evidence of the importance of macroeconomic variables in credit risk management can be found in the literature such as in Bruche and Gonzales-Aguado (2010), Cantor and Varma (2004), Chava et al. (2011), Jankowitsch et al. (2014), Ludvigson and Ng (2009), Mora (2015), and Nazemi and Fabozzi (2018). We have used the database from the Federal Reserve Bank of St. Louis (FRED, Federal Reserve Economic Data) complemented by aggregate default data from Fitch to retrieve 182 macroeconomic variables used in credit risk literature, such as in Acharya et al. (2007), Cantor and Varma (2004), Jankowitsch et al. (2014), Mora (2015), and Nazemi and Fabozzi (2018). The macroeconomic data are retrieved between 2000 (one year before the start of the recovery rate observation period) and 2016. The macroeconomic variables are listed in Appendix A.

The last data source for recovery rate estimation is news from front page articles

of the Wall Street Journal. We merge our dataset with news-based measures of uncertainty reported by Manela and Moreira (2017). The word frequency data yielded by this process is regressed on the volatility index VXO with a support vector machine to generate a news-implied uncertainty measure. Thus, they incorporate a measure of the investors' mood that goes beyond commonly used hard data. The relationship between investors' uncertainty and implied volatility is robust also when controlling for realized stock market volatility. Indeed, their work is based on the premise that news reflects the interest of readers and that the words used by the business press express the concerns of the average investor. They classify the texts to determine the sources of uncertainty within the news by applying commonly used text-analysis methods such as WordNet and WordNet::Similarity. We use the monthly time series data to gauge investors' uncertainty.

We exclude one bond from our analysis, because the data was corrupt. The remaining 2079 bonds exhibit an average recovery rate of 45.57% and a sample standard deviation of 35.04%, as illustrated in Table 4.1. We combine the seniority classes, junior subordinate and subordinate, into one class because these two classes contain the fewest observations. In general, the expectation (senior creditors have the highest recovery rate) regarding the average recovery rates within the seniority classes is met. Subordinated bonds exhibit the lowest average recovery rate at 8.15%, while senior subordinated bonds have the second lowest average recovery rate. Accordingly, senior secured bonds have the highest average recovery rate at 61.91%. Moreover, defaulted bonds from the utility sector have the highest average recovery rate, whereas defaulted bonds from the telecommunications sector have the lowest average recovery rate (71.61% vs. 18.54%).

The histogram of the relative frequency of the observed recovery rates in our sample, presented in Figure 4.2, exhibits two peaks. The class between 0% and 10% contains approximately 640 defaulted bonds. There is another peak in distribution at the class of values between 60% and 70%. However, the observed distribution does not appear to be a bimodal distribution.

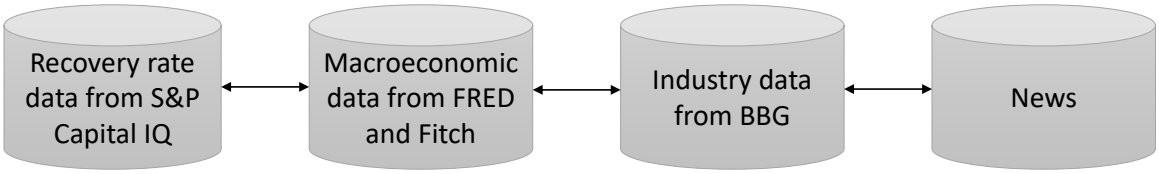

Figure 4.1: Data sources for the target variable and the explanatory variables

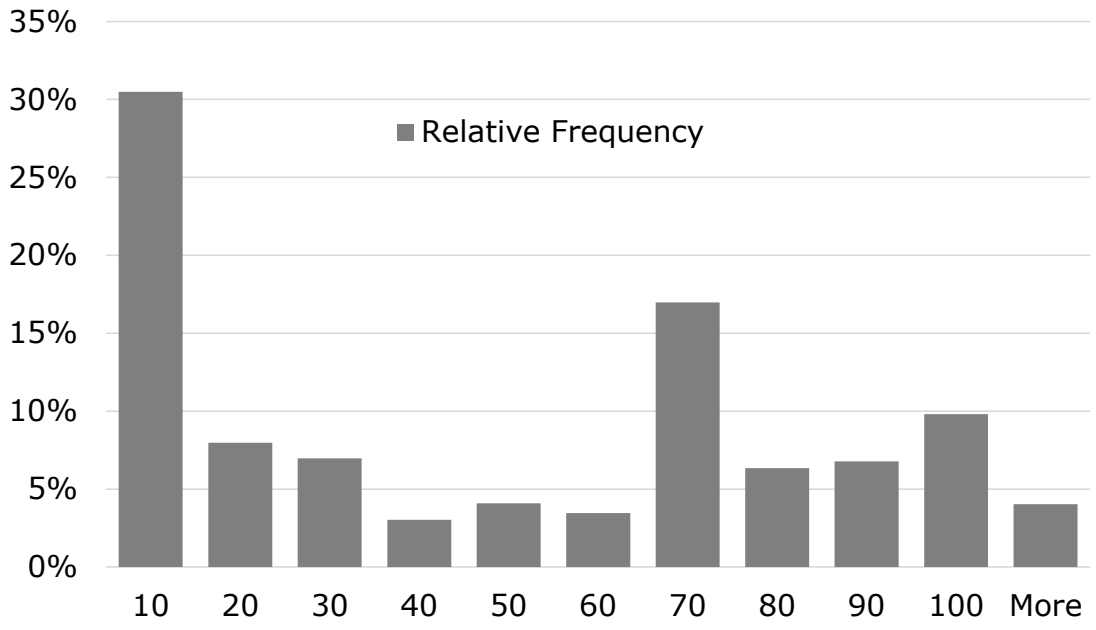

Figure 4.2: Relative frequency of the recovery rates for the defaulted U.S. corporate bonds from 2001 to 2016.

Table 4.1: Descriptive statistics of the recovery rates for all bonds (Panel A), across seniority classes (Panel B) and across industries (Panel C). We report the mean, standard deviation (Std), 10th percentile (p_{10}), first quartile (p_{25}), median, third quartile (p_{75}), 90th percentile (p_{90}), and number of bonds (#).

	Mean	Std	p_{10}	p_{25}	Median	p_{75}	p_{90}	#
Panel A								
All bonds	45.57	35.04	5.00	10.00	43.50	71.96	95.57	2079
Panel B: Recovery rates across seniority								
Senior Unsecured	46.25	34.51	7.50	10.00	48.00	71.00	95.41	1715
Senior Subordinated	24.10	28.34	0.50	2.25	15.50	36.00	72.33	158
Subordinated	8.15	11.98	0.13	0.13	3.00	12.50	18.00	21
Senior Secured	61.91	35.28	5.00	30.00	70.25	94.75	101.15	185
Panel C: Recovery rates across industry								
Utilities	71.61	27.24	38.50	48.25	80.00	94.96	103.08	105
Financials	56.76	34.36	10.00	10.00	67.97	81.95	98.11	1059
Materials	29.95	31.88	1.43	7.50	15.50	48.69	72.58	115
Telecommunication	18.54	24.86	1.04	3.25	10.50	19.50	53.95	124
Consumer Cyclical	31.61	28.31	2.26	5.63	24.56	51.38	74.00	282
Consumer Non-cyclical	39.51	34.78	0.75	7.63	27.00	81.50	82.80	33
Energy	20.08	21.06	1.00	4.66	11.38	29.16	53.75	94
Technology	34.75	33.46	2.00	5.75	23.75	65.31	85.68	74
Industrial	34.69	30.21	3.81	11.00	22.25	48.00	82.50	156

4.5 Empirical analysis of recovery rates' prediction

First, we examine the relation between news and the recovery rate. Second, we investigate the recovery rate estimation in out-of-sample and out-of-time (intertemporal) settings. Finally, we rank the groups of explanatory variables based on their permutation importance for recovery rate prediction.

4.5.1 Analysing the news' impact on recovery rates

Table 4.2 presents an overview of the linear regression specifications based on the entire dataset of 2079 corporate bonds. The recovery rate of the defaulted U.S. corporate bond is the dependent variable. Model (1) incorporates the text-based measures to seniority dummies, industry variables, and bond characteristics as independent variables. In contrast, model (2) considers the seven macroeconomic variables selected through stability selection, in addition to seniority, industry, and bond variables. Finally, we combine the independent variables from models (1) and (2) in model (3).

Adding the text-based measures for uncertainty instead of the selection of seven macroeconomic variables to the basic independent variables yields an improved in-

sample fit (adj. R^2 of 45.36% compared to 44.62%). Combining both groups of variables in (3) generates a further improvement of in-sample fit to an adj. R^2 of 48.26%. We show the significance of three out of five text-based measures of news implied volatility even when controlling for the effects of macroeconomic variables in (3).

We, moreover, add five text-based sources of uncertainty as measures for investors' uncertainty in model (3). Five categories of text can be identified as origins of uncertainty in our analysis: government, intermediation, stock markets, war, and unclassified. We observed that uncertainty-related intermediation has a significantly negative impact in models (1) and (3). The most frequent word counts in the intermediation category are the following: financial, business, bank, credit, and loan. Intermediation-related uncertainty primarily spikes during financial crises and periods of bank failures. Thus, the observed negative impact on recovery rates is in accordance with the intuitive expectation of lower recovery rates during times of financial distress.

News-related uncertainty from the unclassified category has a significantly negative coefficient in both models (1) and (3). The most frequently occurring words in the category unclassified are: U.S., Washington, gold, special, and treasury. The occurrence of the terms 'gold' and 'treasury' indicates macroeconomic uncertainty, as these assets are often regarded as safe havens. Assuming that recovery rates are lower in an environment with increased macroeconomic uncertainty, this interpretation of the category unclassified would explain the significantly negative coefficient of this source of uncertainty.

News-related uncertainty from the government category is the only source of uncertainty that exhibits a significantly positive coefficient in our analysis. The most frequently occurring words in this category are: tax, money, rates, government, and plan. These terms do not necessarily bear a negative connotation. For instance, the prospect of tax cuts or a more expansive fiscal policy might increase uncertainty in news from the government category. Thus, the expectation of positive government policies is one possible explanation for the significantly positive impact on recovery rates in our analysis.

Stock-market-related uncertainty is represented most frequently through the following words: stock, market, stocks, industry, markets. With uncertainty about financial

crises already reflected by highly significant intermediation-related uncertainty, we observed that stock-market-related uncertainty possesses a negative but insignificant coefficient in models (1) and (3). Furthermore, there was minimal variance in war-related uncertainty over our observation period and thus, it is not a significant determinant in the linear regression analysis.

Overall, considering the significance of three out of the five text-based measures, even when controlling for macroeconomic effects in model (3), there appears to be a time-varying influence of investors' mood on recovery rates. Hence, we can conclude that the effect measured by the text-based measures of uncertainty possess additional predictive power for recovery rates, and are not simply mirroring the already known significance of macroeconomic variables for recovery rate prediction.

4.5.2 State-of-the-literature out-of-sample recovery rate prediction

We used two different prediction settings in this study: First, we predicted out-of-sample by randomly stratifying the dataset for the seniority classes. After using a ten-folds cross validation to select the hyperparameters based on the root-mean-squared errors (RMSEs) on the training set (70% of the data), we predicted out-of-sample on the test set (30% of the data). This way, we are able to determine the optimal number of trees and minimum leaf size for the random forest, as well as the cost C and the kernel width γ for the SP LS-SVR. We follow the recommendation from Bui et al. (2017) to use α=0.5 for an MSE loss when applying their power expectation propagation approach.

Table 4.3 demonstrates that machine learning techniques outperform traditional statistical techniques during out-of-sample predictions. Using a random partition of 70% of the dataset as the training set we are able to mitigate the risk of overfitting. In addition to evaluating a wide range of prediction methods, we compare the performance using stability selection, the SparseStep algorithm, and the MC+ algorithm to select the most important macroeconomic variables.

Without regard for the selection technique used to determine the macroeconomic

Table 4.2: This table presents the results of the linear regression specifications. The recovery rate of the respective bond is the independent variable. In (1) we add the news-based measures to seniority dummies, industry variables, and bond characteristics as independent variables. In contrast, in (2) we consider the macroeconomic variables selected by stability selection in addition to the seniority dummies, industry variables, and bond characteristics. In (3) we add the combination of text-based measures and the selection of macroeconomic variables to the base model. The respective t-statistics for each variable are presented in parentheses. Statistical significance at the 99% level is indicated with ***, significance on the 95% level is indicated with ** and significance on the 90% level is marked with *.

Variable	(1)	(2)	(3)
Intercept	38.1582***	46.0907***	33.1257***
	(11.8707)	(18.5324)	(8.0126)
Government	31.7482***		39.6178***
	(10.9206)		(11.7446)
Intermediation	-3.579***		-6.1713***
	(-3.2054)		(-3.9478)
Securities Markets	-0.646		-0.822
	(-0.3849)		(-0.4804)
War	4.4547		-11.011
	(0.5826)		(-1.2071)
Unclassified	-1.392***		-0.547**
	(-6.3147)		(-2.2647)
Manufacturers: Inventories to Sales Ratio		61.4796***	50.183**
		(2.9021)	(2.3043)
Number of Civilians Unemployed for Less Than 5 Weeks		-0.0167***	-0.0148***
		(-3.6623)	(-3.1979)
30-Year Conventional Mortgage Rate		8.2292***	10.0959***
		(5.8475)	(7.0106)
3-Month Commercial Paper Minus Federal Funds Rate		-5.7185**	4.2306
		(-1.9842)	(1.3994)
Light Weight Vehicle Sales: Autos & Light Trucks		-0.1929	1.294*
		(-0.3101)	(1.9188)
Nonfarm Business Sector: Unit Labor Cost		-1.4842***	-1.3715***
		(-3.8899)	(-3.5514)
Trade Weighted U.S. Dollar Index: Major Currencies		-1.1194***	-1.2015***
		(-6.6632)	(-7.1092)
Adj. R^2	0.4536	0.4462	0.4826
RMSE	25.7603	25.9202	25.0247
MAE	19.9213	20.1421	19.2381
AIC	1.95E+04	1.95E+04	1.93E+04
BIC	1.96E+04	1.96E+04	1.95E+04
Number of bonds	2079	2079	2079
Seniority	Yes	Yes	Yes
Industry	Yes	Yes	Yes
Bond Characteristics	Yes	Yes	Yes

variables, all four machine learning techniques (i.e. regression tree, a power expectation propagation approach, SP LS-SVR, and random forest) outperform the two traditional techniques in both performance evaluation metrics, RMSE and MAE. Independent of which selection technique is applied, random forest exhibits the best predictive out-of-sample performance. Moreover, for all six prediction techniques, applying SparseStep for macroeconomic variable selection yields the best predictive accuracy. Thus, in determining that selecting the macroeconomic variables using SparseStep instead of LASSO increases predictive accuracy, we improve upon the study from Nazemi and Fabozzi (2018), which uses LASSO to select the macroeconomic variables. Lastly, the difference between the two remaining selection techniques, MC+ and stability selection, is modest.

The lowest RMSE (20.6838) is observed when selecting the macroeconomic variables with SparseStep and using random forest for prediction. Using SP LS-SVR (20.9890) and the power expectation propagation approach (21.2664) decreases the predictive accuracy slightly. Moreover, regression tree (best RMSE 22.4956) has the lowest predictive power of the machine learning techniques. Among the traditional approaches, inverse Gaussian regression has a minor advantage in predictive capacity compared with linear regression for all three selection techniques. Applying SparseStep for the macroeconomic variables' selection yields the lowest RMSE for linear regression and inverse Gaussian regression. For this reason, we use SparseStep during out-of-time prediction.

Although comparability of the performance measures across datasets is limited, our results for out-of-sample estimation are in accordance with the best results in the literature. For example, the lowest RMSE reported by Yao et al. (2015) is 0.2136 for SP LS-SVR during an out-of-sample prediction study. Moreover, Nazemi and Fabozzi (2018) report the lowest RMSE of 0.1750 for LS-SVR with different intercepts for each seniority class during a ten-folds cross validation. The lowest RMSE during a 12-folds cross validation in a study conducted by Kalotay and Altman (2017) is 0.27 for the regression tree.

In summary, during out-of-sample estimation, all four machine learning techniques outperform the two traditional approaches (i.e. linear regression and inverse Gaussian regression) irrespective of which selection technique is utilized. While this relationship

is documented within literature, such as in Qi and Zhao (2011), Yao et al. (2015), Kalotay and Altman (2017), and Nazemi and Fabozzi (2018), the literature on corporate bonds' recovery rate for out-of-time prediction is sparse. In the following, we address this gap within the literature.

Table 4.3: This table shows the performance measures from out-of-sample prediction on the testing set which is a random partition of the dataset (30%) while the remaining 70% of the dataset were used for training and determining the hyperparameters during cross validation. (SP LS-SVR: Semi-Parametric Least-Squares Support Vector Regression; Lin. Reg.: Linear Regression; Reg. Tree: Regression Tree; PEP: Sparse Gaussian Process Approximation with Power Expectation Propagation; RF: Random Forest; IG: Inverse Gaussian Regression)

	SparseStep		MC+		Stability Selection	
Model	RMSE	MAE	RMSE	MAE	RMSE	MAE
SP LS-SVR	20.9890	13.2027	20.9971	13.5843	21.4105	13.5146
Lin. Reg.	24.8969	18.9199	25.1544	19.2876	25.2331	19.3116
Reg. Tree	22.4956	14.0037	22.5373	14.4637	23.3830	14.8230
PEP	21.2664	14.0712	21.3650	13.8618	21.2667	13.8177
RF	20.6838	13.2145	20.7231	13.2625	21.0394	13.5151
IG reg.	24.0352	17.9865	24.2890	18.1841	24.4879	18.2376

4.5.3 Intertemporal prediction of the recovery rate

In a second step, we predict out-of-time. As outlined by Kalotay and Altman (2017), out-of-time prediction addresses several issues. Kalotay and Altman (2017) argue that considering the likelihood of time variation in recovery rates to report out-of-sample performance on a random split of the dataset is less appropriate. Instead, they emphasize the importance of accounting for time variation in recovery rates. In particular, testing out-of-time performance ensures that only the sample points observed before the default event are used for training. Furthermore, only investigating out-of-time performance prevents data points from the same issuer and the same exposure being a part of both the training and test sets.

We train our models (including the data) until 2011, and use data from the remainder of the sample period (from 2012 to 2016) as a test set. Following Kalotay and Altman (2017), for ease of comparison, we draw a sample of 100 bonds from the test set and calculate the average recovery rate on this sample, weighting the bonds

equally. This procedure is repeated 10,000 times. Moreover, we repeat this analysis over time. Starting with the training set from 2002 to 2011, we add an additional year of data to the training set until we reach the end of the dataset, using training data up to 2014. The bonds from the two years following the training period are used as test set, whereby we sample nine bonds from the respective two-year period and repeat this step 2,000 times.

The out-of-time performance of our models is presented in Table 4.4.[11] The bonds from 2001 to 2011 are used as a training set while the bonds from 2012 to 2016 are used as test set for sampling.

Again, machine learning techniques outperform the traditional approaches for all prediction techniques. In particular, the predictive accuracy of inverse Gaussian regression and linear regression decreases significantly. In contrast to out-of-sample prediction, random forest is the worst performing machine learning technique for out-of-time prediction with an RMSE of 13.5294. The power expectation propagation approach yields the lowest RMSE of 2.6887 while SP LS-SVR (4.2736) and regression tree (5.1717) exhibit slightly lower predictive capacity.

Table 4.5 depicts the out-of-time performance of our models when retraining the models each year. Starting with a training set that includes bonds until 2011, we extend the training set with new bonds each year and use the bonds from the following two years as the test set for sampling. For instance, in the first step, we use the bonds from 2001 to 2011 as the training set and the sample from the 2012-2013 bonds for prediction. In the next iteration, we extend our training set to include the bonds from 2012 and use the bonds from 2013 and 2014 for sampling.

Based on RSME and MAE, the best performing model is the power expectation propagation approach (RMSE of 11.7634), followed by SP LS-SVR (13.1569) and regression tree (13.7023). However, the prediction performance on the quantiles of the recovery rate distribution offers further insight. While the power expectation propagation approach is the best performing model in terms of RMSE and MAE, it has

[11] As we yield the most accurate predictions with SparseStep during out-of-sample prediction, we report only the results obtained from applying SparseStep for macroeconomic variable selection during out-of-time prediction. The results using MC+ and stability selection are consistent with the results reported for SparseStep. These results are not reported here, but are available from the authors.

the lowest percentage deviation among all techniques for only the 1st-, 5th-, and 75th-percentiles. In contrast, regression tree has the lowest percentage deviation for the 10th- (deviating 127.4%) and 25th- (deviating 34.36%) percentiles, while SP LS-SVR has the lowest percentage deviation for the median (5.05%), and random forest has the lowest percentage deviation for the 90th-percentile (deviating -1.56%).

Whereas Kalotay and Altman (2017) report their lowest RMSE (6.8) for out-of-time estimation without retraining for a mixture model with bagging, we report lower RMSEs of 2.7 for the power expectation propagation approach and 4.3 for the SP LS-SVR. The results are similar for out-of-time prediction when retraining the models annually. Each of our three best-performing machine learning techniques (i.e. the power expectation propagation approach (RMSE: 11.8), the SP LS-SVR (RMSE: 13.2), and the regression tree (RMSE: 13.7)) outperform their best-performing technique, i.e. a mixture model with bagging (RMSE: 15.5). Comparing our best techniques with those of Kalotay and Altman (2017), ours outperform theirs for the median and the higher percentiles (75% and 90%), but not for the lower percentiles (1%, 5%, 10%, and 25%).

More traditional approaches, such as linear regression and inverse Gaussian regression, experience significant deterioration during the out-of-time prediction compared with Kalotay and Altman's (2017) out-of-sample performance. In contrast, the predictive accuracy of the machine learning techniques, such as the power expectation propagation approach and SP LS-SVR, does not decline when switching from out-of-sample estimation to out-of-time estimation.

Table 4.6 reports the average performance measures across all time steps, as well as presents the predictive performance for each of the four two-year-ahead sub-periods following the respective period used for training each model. Hence, we are able to demonstrate the consistency of our modeling approaches.

4.5.4 Permutation importance of groups of explanatory variables

In the following, we investigate the permutation importance of each group of variables for the performance of the random forest at recovery rate prediction. We build

Table 4.4: This table shows the performance measures from out-of-time prediction sampling from the testing set (from 2012 to 2016) while the data from 2001 to 2011 are used for training and determining the hyperparameters during cross validation. The SparseStep algorithm is used to select the most informative macroeconomic variables. (SP LS-SVR: Semi-Parametric Least-Squares Support Vector Regression; Lin. Reg.: Linear Regression; Reg. Tree: Regression Tree; PEP: Sparse Gaussian Process Approximation with Power Expectation Propagation; RF: Random Forest; IG: Inverse Gaussian Regression)

		SparseStep					
	Actual	IG reg.	Lin. Reg.	Reg. Tree	SP LS-SVR	PEP	RF
Mean	32.4095	76.7641	78.2838	37.1176	36.2062	34.1537	45.7951
Std	2.1677	1.3181	1.5030	1.6219	0.7990	0.9622	0.8272
1%	27.4195	73.7258	74.8435	33.3696	34.3637	31.9405	43.9059
		168.88%	172.96%	21.70%	25.33%	16.49%	60.13%
5%	28.8675	74.6184	75.8140	34.4713	34.9157	32.5985	44.4500
		158.49%	162.63%	19.41%	20.95%	12.92%	53.98%
10%	29.6229	75.0592	76.3542	35.0428	35.1853	32.9160	44.7320
		153.38%	157.75%	18.30%	18.78%	11.12%	51.01%
25%	30.9674	75.8629	77.2716	36.0258	35.6595	33.4930	45.2336
		144.98%	149.53%	16.33%	15.15%	8.16%	46.07%
50%	32.4085	76.7598	78.2634	37.1102	36.2066	34.1568	45.7914
		136.85%	141.49%	14.51%	11.72%	5.39%	41.29%
75%	33.8399	77.6559	79.2965	38.2048	36.7370	34.8010	46.3507
		129.48%	134.33%	12.90%	8.56%	2.84%	36.97%
90%	35.2057	78.4484	80.2251	39.1859	37.2386	35.3864	46.8660
		122.83%	127.88%	11.31%	5.77%	0.51%	33.12%
RMSE		44.4119	45.9333	5.1717	4.2736	2.6887	13.5294
MAE		44.3545	45.8743	4.7300	3.8397	2.1923	13.3856

Table 4.5: This table shows the performance measures from out-of-time prediction when all models are retrained every year. Starting with a training set including bonds until 2011 we extend the training set with new bonds each year and use the bonds from the following two years as test set. So, in the first step we use the bonds from 2001 to 2011 as training set and sample from the bonds from 2012 and 2013. In the next iteration, we extend our training set to include the bonds from 2012 and use the bonds from 2013 and 2014 as test set. The SparseStep algorithm is used to select the most informative macroeconomic variables. (SP LS-SVR: Semi-Parametric Least-Squares Support Vector Regression; Lin. Reg.: Linear Regression; Reg. Tree: Regression Tree; PEP: Sparse Gaussian Process Approximation with Power Expectation Propagation; RF: Random Forest; IG: Inverse Gaussian Regression)

		SparseStep					
	Actual	IG reg.	Lin. Reg.	Reg. Tree	SP LS-SVR	PEP	RF
Mean	35.4642	65.1487	64.5576	42.6204	37.5660	40.493	46.4706
Std	13.2443	6.5806	6.8764	11.2331	4.1321	8.16991	4.1480
1%	8.7095	49.2976	48.9102	19.8052	28.2487	17.6113	37.4106
		466.02%	461.57%	127.40%	224.34%	102.21%	329.54%
5%	13.8126	53.8581	53.2014	25.3224	30.9192	24.9364	40.0503
		289.92%	285.17%	83.33%	123.85%	80.53%	189.96%
10%	17.3750	56.3814	55.5540	28.7439	32.4446	29.5408	41.3762
		224.50%	219.74%	65.43%	86.73%	70.02%	138.14%
25%	25.7049	60.8349	59.8974	34.5368	34.9132	36.1703	43.6340
		136.67%	133.02%	34.36%	35.82%	40.71%	69.75%
50%	35.8299	65.3895	64.6008	41.6541	37.6385	41.4314	46.3077
		82.50%	80.30%	16.26%	5.05%	15.63%	29.24%
75%	44.5556	69.8571	69.2250	50.2680	40.3493	45.8537	49.1103
		56.79%	55.37%	12.82%	-9.44%	2.91%	10.22%
90%	52.6980	73.5200	73.5126	57.9349	43.0418	49.7292	51.8733
		39.51%	39.50%	9.94%	-18.32%	-5.63%	-1.56%
RMSE		33.0883	32.6534	13.7023	13.1569	11.7634	17.2256
MAE		29.7743	29.2172	11.0205	10.6562	9.41088	13.9478

Table 4.6: This table shows the performance measures for each two-year ahead subperiod from out-of-time prediction when all models are retrained every year. The first column shows the last year that is included in the training set. # of bonds denotes the number of bonds in each two-year ahead period which is used as test set for sampling. Starting with a training set including bonds until 2011 we extend the training set with new bonds each year and use the observations from the following two years as test set. So, in the first step we use the observations from 2001 to 2011 as training set and the bonds from 2012 and 2013 as test set. In the next iteration, we extend our training set to include the bonds from 2012 and use the bonds from 2013 and 2014 as test set. The SparseStep algorithm is used to select the most informative macroeconomic variables. (SP LS-SVR: Semi-Parametric Least-Squares Support Vector Regression; Lin. Reg.: Linear Regression; Reg. Tree: Regression Tree; PEP: Sparse Gaussian Process Approximation with Power Expectation Propagation; RF: Random Forest; IG: Inverse Gaussian Regression)

			SparseStep					
# of bonds			IG reg.	Lin. Reg.	Reg. Tree	SP LS-SVR	PEP	RF
2011	30	RMSE	28.4971	29.8182	9.2928	7.8582	8.4563	9.3318
		MAE	27.6984	28.9866	7.3983	6.3000	6.8337	7.6770
2012	32	RMSE	17.8793	16.8617	12.3584	12.1457	9.5560	9.6305
		MAE	15.6871	14.6412	9.9761	9.7989	7.6453	7.7579
2013	59	RMSE	30.3518	28.3460	16.6306	9.5900	11.4688	15.9382
		MAE	28.1792	26.1179	13.5029	7.6879	9.2671	13.6061
2014	45	RMSE	48.2324	47.8334	15.3412	19.7783	16.0980	27.4414
		MAE	47.5327	47.1232	13.2046	18.8379	13.8974	26.7500

11 groups of independent variables, as detailed in Appendix A: industry, bond characteristics, seniority, news, and the macroeconomic variables which are separated into the following groups: financial conditions, micro-level factors, business cycle, monetary measures, corporate profitability (on a macro level), international competitiveness, and stock market. We scale the permutation importance of each group such that the importance of the most important group of variables equals 100. We subsequently examine the importance ranking of the groups of variables for the U.S. corporate bonds that defaulted between 2001 and 2016.

As illustrated in Table 4.7 and Figure 4.3, bond characteristics are the most important group of variables for recovery rate prediction in our analysis. Thus, the significance of bond characteristics determined by Jankowitsch et al. (2014) is further confirmed in our study. The importance of the seniority of the defaulted bond (ranked second, 30.7945) is in accordance with the significance of the seniority reported in the literature, such as in Cantor and Varma (2004) and Jankowitsch et al. (2014). Moreover, the importance of stock market indicators (ranked third, 14.3448) confirms the significance of the return on the market index, as reported by Cantor and Varma (2004).

Interestingly, the group of news variables is ranked higher than the industry variables (9.5908 compared with 6.1447). The literature has paid little attention to variables that indicate international competitiveness, which rank fourth in our analysis with an importance of 13.2206. Having an importance of 2.9321, business cycle variables, including regularly used variables such as GDP growth and the unemployment rate (e.g. by Yao et al. (2015)), rank second to last in our analysis.

Micro-level factors such as the fed funds rate and the term structure which were reported to be significant by Jankowitsch et al. (2014), and were also considered by Qi and Zhao (2011), are ranked seventh in our analysis with an importance of 4.5070. However, among the macroeconomic variables, micro-level factors constitute the group with the third-highest rank. Financial conditions and monetary measures have not been investigated in the literature, but are also not important in our ranking for the entire dataset (3.4836 and 2.2154, respectively). Ours is the first study to examine news variables as independent variables for recovery rate prediction.

The industry of the defaulted bond has been found to be an important determinant

of recovery rates by Altman and Kishore (1996). Furthermore, Acharya et al. (2007) have introduced two industry distress dummy variables indicating negative sales growth of the respective industry and a performance of the industry index worse than - 30% in the preceding year. These industry distress dummy variables are part of the industry group in our analysis. In our analysis however, industry variables have an importance of 6.1447 and rank only sixth.

Table 4.7: Ranking groups of variables by permutation importance for all defaulted bonds from 2001 to 2016

Rank	Entire dataset	Importance
1	Bond Characteristics	100.0000
2	Seniorities	30.7945
3	Stock Market Indicators	14.3448
4	International Competitiveness	13.2206
5	News	9.5908
6	Industry	6.1447
7	Micro-Level Factors	4.5070
8	Corporate Profitability (Macro)	4.3065
9	Financial Conditions	3.4836
10	Business Cycle	2.9321
11	Monetary Measures	2.3154

4.6 Macroeconomic stress testing

4.6.1 Motivation and literature

To the best of our knowledge, there do not exist any studies modeling recovery rates under macroeconomic stress. At best, studies make simplifying assumptions regarding the recovery rates under macroeconomic stress or use other simplifying heuristics for modeling the recovery rate part in a credit risk stress test. As a constant recovery rate, as it is assumed by Virolainen (2004), seems rather unrealistic in a situation of macroeconomic stress, there might be interesting implications for the behavior of the recovery rate under stress.

There are several studies stress testing default probabilities that provide insight for

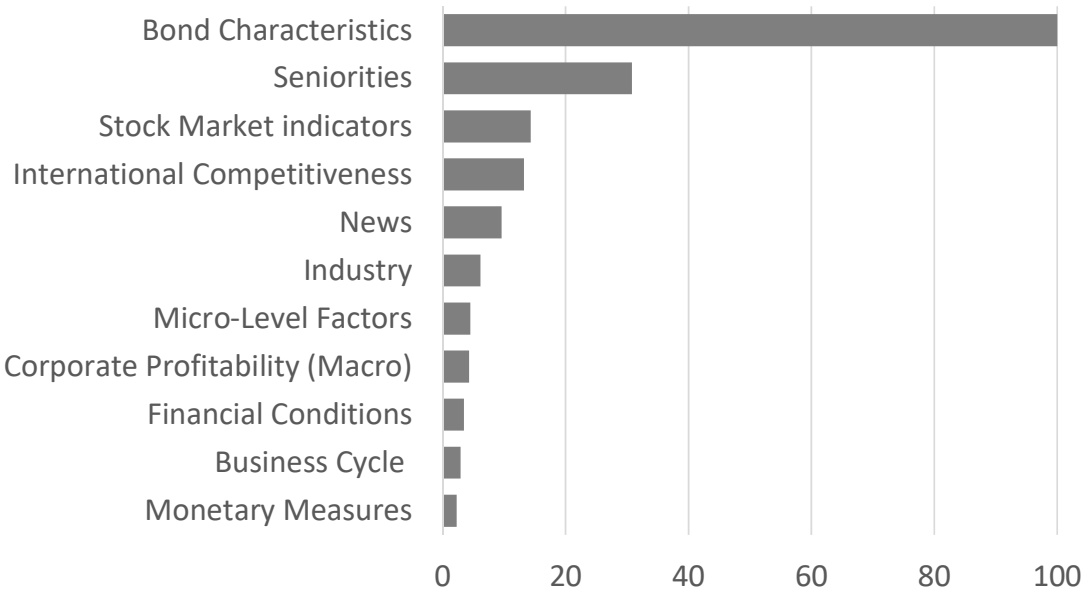

Figure 4.3: Ranking of the permutation importance of the groups of independent variables for the full period from 2001 to 2016, scaled such that the biggest importance equals 100.

our analysis. Bellotti and Crook (2013a) include macroeconomic variables in a survival analysis model for the probability of default in a large set of credit card accounts. They show that the addition of macroeconomic covariates as time series increases the model fit and yields a modest improvement in predictions of default on an out-of-sample test set. They also state that this model is suitable for stress tests. In a later study, Bellotti and Crook (2013b) present discrete time survival models of default rates for credit cards incorporating behavioral data about the credit card holders as well as macroeconomic conditions during the credit card lifetime. They perform the stress tests applying Monte Carlo simulation on the Cholesky decomposition of the macroeconomic variables. They conclude that the Value-at-Risk from their stress test result seems to be on average in line with other studies. Further, Bellotti and Crook (2013b) employ a logistic survival model to estimate a dynamic model of default for retail credit data. For the derivation of the macroeconomic factors they extract the principal components of the macroeconomic input variables. To demonstrate the influence of the macroeconomic factors (PCs) on their model they apply Monte Carlo Simulation to simulate the macroeconomic factors used as input for their PD estimation model. These macroeconomic factors are drawn from the historical distribution of the respective factors. The main finding from Bellotti and Crook (2013b) is that their proposed methodology is sufficient to produce realistic looking loss distributions, that is loss distributions with a long tail.

4.6.2 Value at Risk models and stress testing

According to Alexander and Sheedy (2008) the empirical Value-at-Risk (VaR) is very popular in the industry. Apart from the assumption that returns are independent and identically distributed it makes no assumptions about the distribution of past returns. VaR_α is the loss that should not be exceeded in more than α-% of the cases. So, following Bellotti and Crook (2013b) VaR can be calculated as the average recovery rate that is predicted at the α-percentile of simulated macroeconomic data, denoted as macroeconomic data_α:

$$\text{VaR}_\alpha = \frac{1}{N}\sum_{i=1}^{N}(R_i(\cdot,\cdot,\text{macroeconomic data}_\alpha)) \tag{4.8}$$

where R_i denotes the respective regression model for prediction of sample i: LS-SVR, SP-SVR or linear regression. The macroeconomic data are the macroeconomic variables selected by LASSO. We standard normalize each macroeconomic data series. For

performing the stress test we draw 10000 random values from the standard normal distribution for each macroeconomic model component as we observe convergence of the Monte Carlo Simulation for this number of draws. The Conditional Value-at-Risk (CVaR), which is also known as expected tail loss, is defined as the expected loss conditional on exceeding the VaR:

$$\text{CVaR}_\alpha = \frac{1}{N \times \alpha \times 10000} \sum_{i=1}^{N} \sum_{a=1}^{\alpha \times 10000} (R_i(\cdot, \cdot, \text{macroeconomic data}_a)) \tag{4.9}$$

where a denotes the index passing through all samples of simulated macroeconomic data below the α-quantile. That is, the expected value of all empirical recovery rates below VaR. In particular, it is important to notice that the key metric for evaluating the applicability of our method is different in stress tests from ordinary forecasts because one prefers erring on the side of caution during a stress test while during regular forecasting accuracy measures such as mean squared error or mean absolute error are used. The 5%-quantile and the 10%-quantile of the simulated macroeconomic model components are applied as macroeconomic shocks. We have controlled for negative influences of variables in a linear regression model and use the 90%-quantile respectively the 95%-quantile for the respective macroeconomic data. As performance measures of our analysis we examine the Value at Risk (VaR) and the conditional VaR (CVaR) of the recovery rates.

4.6.3 Comparison of our stress testing results

We determine the hyperparameters of the LS-SVR and the semiparametric LS-SVR in a five-folds cross validation on the training set. We extract the principal components of the 182 macroeconomic variables in our training set. We include the seven first principal components so that more than 95% of the variance among the 182 macroeconomic variables is explained. LASSO selects 26 non-zero loadings within the training set, while we allow for a maximum of 30 non-zero loadings. All calculations for this analysis are conducted in MATLAB.

As can be seen in Table 4.8 for both methods of extracting the macroeconomic information SP LS-SVR predicts higher recovery rates than LS-SVR. The lowest predictions of the recovery rate under stress are yielded by the models including the first seven principal components of all macroeconomic variables. In contrast, the models

Table 4.8: Comparing the stress measures from semiparametric LS-SVR, LS-SVR and linear regression. We vary the extraction method for the macroeconomic factors between PCA and LASSO. The most adverse conditions from the simulated principal components are included at the 0.1-quantile respectively the 0.05-quantile. If the factor has a negative coefficient, when regressed on the recovery rate within the training set, the 0.9-quantile respectively the 0.95-quantile are used.

RR	PCA			LASSO		
	SP-SVR	LS-SVR	linReg	SP-SVR	LS-SVR	linReg
$\text{VaR}_{0.1}$	12.0812	4.1172	15.0167	18.9258	10.4730	17.4870
$\text{CVaR}_{0.1}$	7.5695	0.8418	8.9156	10.2388	2.9559	8.9948
	SP-SVR	LS-SVR	linReg	SP-SVR	LS-SVR	linReg
$\text{VaR}_{0.05}$	6.6261	0.1355	9.3791	9.1339	1.7119	9.2371
$\text{CVaR}_{0.05}$	6.1365	0.1008	5.5118	6.5550	0.4772	4.6298

using LASSO to select the macroeconomic variables generate predictions under macroeconomic stress in a reasonable corridor.

Under the stress testing rules and capital plan rules of the Dodd-Frank Act, the Board of Governors of the Federal Reserve System (Board of the Fed) is required to conduct an annual stress test for large bank holding companies and other non-bank financial companies determined by the Financial Stability Oversight Council for Federal Reserve supervision (Federal Reserve, 2012, 2014a, 2014b, 2016). For this purpose, each year a baseline scenario, an adverse scenario, and a severely adverse scenario are defined by the Federal Reserve.

We train our models using the bond-specific variables along with the macroeconomic variables selected by the Board of the Federal Reserve. For the stress scenario, we use the severely adverse stress scenarios as outlined by the Board of the Federal Reserve in their supervisory scenarios report at the beginning of each year. The variables with their respective values in the severely adverse stress scenario are presented in Table 4.9. For the calculations of VaR and CVaR we substitute the $\alpha\%$-quantiles of the simulated macroeconomic data with the values of the macroeconomic data under the severely adverse stress scenarios. Due to the fact that it is difficult to find a sensible performance benchmark for macroeconomic stress tests we compare our results inducing simulated macroeconomic stress with our results using the stress scenarios of the Federal Reserve.

The results of the stress tests implemented as described above are presented in Table 4.10. Linear regression is not able to handle the macroeconomic stress well as it predicts a VaR of 83.05% which is too high compared with the out-of-sample average recovery rate of 32.57% without macroeconomic stress. The VaR of 7.61% from the SP LS-SVR using the variables and scenarios defined by the Federal Reserve is lower than all VaR measures but the $\text{VaR}_{0.05}$ from SP LS-SVR with PCA following our approach simulating macroeconomic stress. Overall, the VaR values calculated by using the Fed variables and scenarios compare well to the values calculated by our approach with simulated macroeconomic stress.

4.7 Conclusions

The recovery rate is a key risk parameter in credit risk. Although a substantial amount of literature examines out-of-sample recovery rate estimation for corporate bonds, the majority of approaches suffer from two primary shortcomings: First, the assumption of a time invariant recovery rate distribution is unrealistic. Second, assuming the independence of a sample when multiple defaulted bonds from the same issuer are part of both the training and test sets creates unrealistically accurate predictions. Therefore, it is essential to examine the estimation of this risk factor for defaulted U.S. corporate bonds in an intertemporal setting.

In this study, we investigate the recovery rate prediction of defaulted corporate bonds between 2001 and 2016 in an intertemporal set-up to address these issues. We find that machine learning techniques outperform traditional approaches, such as inverse Gaussian regression and linear regression, during out-of-time predictions. In particular, employing a semiparametric least-squares support vector regression, a power expectation propagation approach, regression tree, or random forest yields significantly higher predictive out-of-time accuracy than the statistical techniques.

We use news-based measures that are developed based on a text-based analysis of news to examine the relationship between the news and the recovery rates of bonds. Interestingly, we find that investors' uncertainty about the government, intermediation, and the economy are significant drivers of recovery rates.

A broad range of macroeconomic variables are included in our analysis, and we

Table 4.9: The national macroeconomic variables and their respective growth values in the severly adverse stress scenario as defined by the Federal Reserve.

Year	Q	δ Real GDP	δ Nom GDP	δ Real Disp Inc	δ Nom Disp Inc	Unempl rate	CPI infl rate	3m t-bill	5y t-note	10y t-bond	BBB corp	Mortg 30 yr	Prime rate	Dow vol	S&P Ret	δ HPI
2013	Q1	-6.1	-4.7	-6.7	-5.9	10	1.4	0.1	NS	1.2	6.4	4.5	NS	76.6	-20.3	-2.6
2013	Q2	-4.4	-3.3	-4.6	-4	10.7	1.1	0.1	NS	1.2	6.7	4.7	NS	76.4	-6.4	-3.1
2013	Q3	-4.2	-3.6	-3.2	-2.8	11.5	1	0.1	NS	1.2	6.8	4.8	NS	79.4	-19.5	-3.4
2013	Q4	-1.2	-1.2	-1.5	-1.8	11.9	0.3	0.1	NS	1.2	6.5	4.7	NS	71.7	-0.7	-3.3
2014	Q1	-6.1	-4	-2.4	-1.9	9.2	0.4	0.1	0.6	1	5.8	4.4	3.3	61.3	-12.4	-3.3
2014	Q2	-3.2	-1.9	0.1	0.8	9.9	0.8	0.1	0.6	1.1	6.1	4.4	3.3	65.7	-14.3	-3.9
2014	Q3	-4	-2.6	-1.1	-0.2	10.7	0.8	0.1	0.6	1.1	6.2	4.4	3.3	57.9	-8.5	-4.3
2014	Q4	-1.5	-0.3	-0.5	0.5	11.1	1.1	0.1	0.6	1.3	6.1	4.4	3.3	42.1	7.5	-4.2
2015	Q1	2.9	4.2	2.9	4.4	5.8	1.9	0.1	2	2.9	4.6	4.5	3.3	18.6	1.2	0.6
2015	Q2	2.9	4.5	2.7	4.3	5.7	2	0.3	2.3	3.1	4.8	4.7	3.4	19.2	1.2	0.6
2015	Q3	2.9	4.7	2.7	4.4	5.6	2.1	0.6	2.5	3.3	5	4.9	3.7	19.6	1.3	0.6
2015	Q4	2.9	4.9	2.8	4.6	5.4	2.1	0.9	2.7	3.5	5.1	5.1	4	19.5	1.3	0.7
2016	Q1	-5.1	-2.6	-0.5	-0.4	6	0.2	0	0	0.2	4.8	3.2	3.3	73.3	-20.2	-2.3
2016	Q2	-7.5	-6.1	-4.1	-3.2	7.2	0.9	-0.2	0	0.4	5.6	3.7	2.9	61.1	-21.3	-3.0
2016	Q3	-5.9	-4.5	-4.5	-3.5	8.3	1.1	-0.5	0	0.4	6	3.9	2.6	67.1	-13.5	-3.5
2016	Q4	-4.2	-2.9	-3.6	-2.5	9.1	1.3	-0.5	0	0.6	6.4	4.1	2.6	59.1	-9.4	-3.9

Table 4.10: The predictions of all three models under the severely adverse stress scenarios as defined by the Board of Governors of the Federal Reserve System. We define the VaR as the average prediction under the severely adverse stress scenario.

RR	SP-SVR	LS-SVR	linReg
VaR	7.6130	11.2718	83.0530

investigate three techniques for the selection of the most informative macroeconomic variables. We find that selecting the macroeconomic variables with innovative machine learning techniques, such as the SparseStep algorithm, yields a modest improvement in the predictive performance of our models. Lastly, in regard to the permutation importance of the groups of macroeconomic variables, we find that bond characteristics, seniority dummy variables, and stock market indicators are the most important groups of variables for corporate bonds' recovery rate prediction.

Chapter 5

Conclusion

The Basel II/III accords allow for an internal calculation of the credit risk parameters to estimate the risk-weighted assets. Thus, the importance of forecasting the credit risk parameters, including probability of default, exposure at default and recovery rates, has increased. As recovery rates have undergone relatively less examination, the improvement of corporate bond recovery rate prediction is the focus of this dissertation.

In Chapter 2 we compare the predictive performance of machine learning techniques, such as the three least-squares support vector techniques, ϵ-insensitive support vector regression and the regression tree approach, with the more traditional linear regression approach. A semiparametric least-squares support vector regression model which utilizes seniority dummies as linear inputs increases the predictive accuracy compared with the standard least-squares support vector regression model and linear regression. In our analysis we also consider macroeconomic variables which are not commonly used in the literature such as housing starts, orders of capitals goods, and stock market volatility.

Compared to recovery rate models in the credit risk literature that include only a few macroeconomic variables, we demonstrate that the addition of a high-dimensional array of 104 macroeconomic variables increases the predictive power of our models. We compare different data reduction techniques for use with these 104 macroeconomic variables, including PCA, kernel PCA, sparse PCA, nonlinear PCA and gradient boosting. In doing so, we show that applying sparse PCA not only allows for better interpretability of the principal components, but also increases the models' predictive capacity. Ranking the macroeconomic variables using gradient boosting reveals that the credit

spread of corporate bonds, the yields offered on corporate bonds, the annual return of the Russell 2000 and the number of unemployed, are the macroeconomic variables with the highest relative importance. Adding the 20 most informative macroeconomic variables from the gradient boosting analysis improves the predictive performance of those models which are easy to interpret, such as the linear regression and the regression tree.

In Chapter 3 we introduce fuzzy decision fusion models for corporate bond recovery rate prediction in a comparative study with four types of support vector regression techniques, a linear regression model and a regression tree. For the creation of the fuzzy rule base we apply the differential evolution algorithm. We show that fuzzy rule-based models outperform the models discussed in the literature for corporate bond recovery rate prediction. Adding the principal components of 104 macroeconomic variables increases the predictive capacity of our models. Further, the Box-Cox transformation of the macroeconomic variables is tested. This transformation does not result in an improved predictive capacity of the fuzzy techniques even though it yields an improvement in the predictive performance of the regression techniques.

In Chapter 4 we compare machine learning techniques for recovery rate prediction with statistical approaches, such as inverse Gaussian regression and linear regression, in an intertemporal set-up. We show that machine learning techniques, such as the regression tree, semiparametric least-squares support vector regression, a power expectation propagation approach and random forest, outperform more traditional approaches during out-of-time prediction. We consider news-based variables, which are developed based on the analysis of text, as explanatory variables for the recovery rate prediction. In our analysis, we find that three out of five news-based measures are significant determinants of recovery rates in our analysis.

We compare stability selection, the SparseStep algorithm and the MC+ algorithm for selecting the macroeconomic variables. We demonstrate that selecting the macroeconomic variables with the SparseStep algorithm produces a modest improvement in predictive accuracy, irrespective of the prediction technique applied. Investigating the permutation importance of each group of variables in our high-dimensional analysis we find that bond characteristics, seniority dummies and stock market indicators are the most important groups of variables for recovery rate prediction. Examining the behaviour of recovery rates under macroeconomic stress we find that, while semipara-

metric least-squares support vector regression yields reasonably low predictions, linear regression produces predictions that are unrealistically high under the severely adverse stress scenario defined by the Federal Reserve.

There remain several interesting research questions that are beyond the scope of this dissertation:

What is the optimal modeling approach for the integration of the forecasts for the probability of default and the recovery rate?

Altman et al. (2005) state that there is a strong link between the market default rate and the aggregate recovery rate. Kalotay and Altman (2017) emphasize the correlation between the probability of default and recovery rates and outline a first approach to modeling the probability of default and the recovery rate at the same time. It is worthwhile investigating whether more sophisticated techniques are applicable in the context of this integrated modeling approach.

Are point estimates sufficient for regulatory and valuation purposes?

This thesis deals with the forecast of recovery rates exclusively in terms of point estimates. For example, approximating sparse Gaussian processes with power expectation propagation allows to generate confidence intervals for the point predictions. Distributional predictions might show substantial value for both risk and trading applications as well as stress testing.

Does the heterogeneity of industry sectors have an impact on the recovery rate beyond the influence of industry dummy and industry distress variables?

Taking into account the heterogeneity of industry sectors, building sector-specific models and including explanatory variables for certain industries might add predictive power to existing approaches, including investigation of: the return of commodities such as oil and natural gas for the energy sector, the term structure and variables measuring alternative central bank operations for the financial sector, durable goods orders for the manufacturing sector and phone part orders for the telecommunications sector. Moreover, including recent industry sector recovery rates in place of the dummy vari-

ables would enable accounting for time variation and structural changes within a sector.

Are there any other drivers of recovery rates?

The use of alternative financial data could help to increase the predictive capacity of machine learning techniques for recovery rate prediction. The examination of rating changes might allow us to differentiate between companies operating in a systematically risky environment and companies that have experienced downward credit migration. The degree of industry segmentation might be helpful to the estimation of buying interest in a defaulted company. Considering the entire U.S. economy as a network of inter-industry customer and supplier relationships, a centrality measure for the position of the defaulted company in the network might provide additional insights. Corporate financial data such as the default barrier, the long term debt ratio and the profitability ratio could add further predictive capacity to the techniques presented in this dissertation. However, recovery datasets with a sufficient size are hard to find. Adding the corporate data mentioned above would have decreased the size of our dataset from 2079 bonds to 506 bonds.

Bibliography

Acharya, V. V., Bharath, S. T., and Srinivasan, A. (2007). Does industry-wide distress affect defaulted firms? Evidence from creditor recoveries. *Journal of Financial Economics*, 85(1):787–821.

Aizerman, A., Braverman, E. M., and Rozoner, L. (1964). Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, 25:821–837.

Alexander, C., Sheedy, E. (2008). Developing a stress testing framework based on market risk models. *Journal of Banking & Finance*, 32(10):2220–2236.

Altman, E. I., Brady, B., Resti, A., and Sironi, A. (2005). The link between default and recovery rates: Theory, empirical evidence, and implications. *Journal of Business*, 78(6):2203–2228.

Altman, E. I. and Kalotay, E. A.(2014). Ultimate recovery mixtures. *Journal of Banking & Finance*, 40:116–129.

Altman, E. I. and Kishore, V. M. (1996). Almost everything you wanted to know about recoveries on defaulted bonds. *Financial Analysts Journal*, 52(6):57–64.

Altmann, A., Toloşi, L., Sander, O., Lengauer, T. (2010). Permutation importance: A corrected feature importance measure. *Bioinformatics*, 26(10):1340–1347.

Basel Committee on Banking Supervision (2006). Basel II: International convergence of capital measurement and capital standards: A revised framework.

Bastos, J. A.. (2010). Forecasting bank loans loss-given-default. *Journal of Banking & Finance*, 34(10):2510–2517.

Bastos, J. A. (2014). Ensemble predictions of recovery rates. *Journal of Financial Services Research*, 46(2):177–193.

Bellotti, T. and Crook, J. (2009). Support vector machines for credit scoring and discovery of significant features. *Expert Systems with Applications*, 36(2):3302–3308.

Bellotti, T. and Crook, J. (2012). Loss given default models incorporating macroeconomic variables for credit cards. *International Journal of Forecasting*, 28(1):171–182.

Bellotti, T., Crook, J. (2013a). Forecasting and stress testing credit card default using dynamic models. *International Journal of Forecasting*, 29(4):563–574.

Bellotti, T., Crook, J. (2013b). Retail credit stress testing using a discrete hazard model with macroeconomic factors. *Journal of the Operational Research Society*, 65(3):340–350.

Board of Governors of the Federal Reserve System (2012). Supervisory scenarios for annual stress tests required under the dodd-frank act stress testing rules and the capital plan rule. Technical report.

Board of Governors of the Federal Reserve System (2014a). Supervisory scenarios for annual stress tests required under the dodd-frank act stress testing rules and the capital plan rule. Technical report.

Board of Governors of the Federal Reserve System (2014b). Dodd-frank act stress test 2014: Supervisory stress test methodology and results. Technical report.

Board of Governors of the Federal Reserve System (2016). Supervisory scenarios for annual stress tests required under the dodd-frank act stress testing rules and the capital plan rule. Technical report.

Box, G. E., Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, S. 211–252.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and Regression Trees*. CRC press.

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

Britto, A. S., Sabourin, R., Oliveira, L. E. (2014). Dynamic selection of classifiers—a comprehensive review. *Pattern Recognition*, 47(11):3665–3680.

Bruche, M., Gonzalez-Aguado, C. (2010). Recovery rates, default probabilities, and the credit cycle. *Journal of Banking & Finance*, 34(4):754–764.

Bui, T. D., Yan, J., Turner, R. E. (2017). A unifying framework for Gaussian process pseudo-point approximations using power expectation propagation. *Journal of Machine Learning Research*, 18(1):3649–3720.

Burduk, R., Walkowiak, K. (2015). Static classifier selection with interval weights of base classifiers. *Intelligent Information and Database Systems*, S. 494–502. Springer.

Cantor, R. and Varma, P. (2004). Determinants of recovery rates on defaulted bonds and loans for North American corporate issuers: 1983-2003. *Journal of Fixed Income*, 14(4):29–44.

Calabrese, R. and Zenga, M. (2010). Bank loan recovery rates: Measuring and nonparametric density estimation. *Journal of Banking & Finance*, 34(5):903–911.

Chalup, S. K. and Mitschele, A. (2008). Kernel methods in finance. In Seese, D., Weinhardt, C., and Schlottmann, F., editors, *Handbook on information technology in finance*. Springer, 655–687.

Chang, C.-C. and Lin, C.-J. (2011). Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

Chava, S., Stefanescu, C., Turnbull, S. (2011). Modeling the loss distribution. *Management Science*, 57(7):1267–1287.

Chen, H. (2010). Macroeconomic conditions and the puzzles of credit spreads and capital structure. *Journal of Finance*, 65(6):2171–2212.

Cheng, D., Cirillo, P. (2018). A reinforced urn process modeling of recovery rates and recovery times. *Journal of Banking & Finance*, 96:1–17.

Danenas, P. and Garsva, G. (2015). Selection of support vector machines based classifiers for credit risk domain. *Expert Systems with Applications*, 42(6):3194–3204.

Dos Santos, E. M., Sabourin, R., Maupin, P. (2008). A dynamic overproduce-and-choose strategy for the selection of classifier ensembles. *Pattern Recognition*, 41(10):2993–3009.

Duffie, D., Eckner, A., Horel, G., and Saita, L. (2009). Frailty correlated default. *Journal of Finance*, 64(5):2089–2123.

Friedman, J. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 1189–1232.

Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., Walther, A. (2017). Predictably unequal? The effects of machine learning on credit markets. *Working Paper*.

Gacto, M. J., Alcalá, R., Herrera, F. (2011). Interpretability of linguistic fuzzy rule-based systems: An overview of interpretability measures. *Information Sciences*, 181(20):4340–4360.

Giannone, D., Lenza, M., Primiceri, G. E. (2017). Economic predictions with big data: The illusion of sparsity. *Working Paper*.

Gregorutti, B., Michel, B., Saint-Pierre, P. (2015). Grouped variable importance with random forests and application to multiple functional data analysis. *Computational Statistics & Data Analysis*, 90:15–35.

Gu, S., Kelly, B. T., Xiu, D. (2018). Empirical asset pricing via machine learning. *Technical Report, University of Chicago*.

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

Hartmann-Wendels, T., Miller, P., Töws, E. (2014). Loss given default for leasing: Parametric and nonparametric estimations. *Journal of Banking & Finance*, 40:364–375.

Hoechstoetter, M. and Nazemi, A. (2013). Analysis of loss given default. *Investment Management and Financial Innovations*, 10(4):70–79.

Hsieh, W. (2007). Nonlinear principal component analysis of noisy data. *Neural Networks*, 20(4):434–443.

Hsieh, W. (2006). Neuralnets for multivariate and time series analysis (neumatsa): a user manual.

Hsu, C.-W., Chang, C.-C., Lin, C.-J.(2003). A practical guide to support vector classification. *Technical report, Department of Computer Science, National Taiwan University*.

Ishibuchi, H., Nozaki, K., Tanaka, H. (1992). Distributed representation of fuzzy rules and its application to pattern classification. *Fuzzy Sets and Systems*, 52(1):21–32.

Jacobs, M., Karagozoglu, A. K. (2011). Modeling ultimate loss given default on corporate debt. *Journal of Fixed Income*, 21(1):6–20.

Jankowitsch, R., Nagler, F., and Subrahmanyam, M. G. (2014). The determinants of recovery rates in the us corporate bond market. *Journal of Financial Economics*, 114(1):155–177.

Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794.

Jurek, A., Bi, Y., Wu, S., Nugent, C. (2014). A survey of commonly used ensemble-based classification techniques. *Knowledge Engineering Review*, 29(05):551–581.

Kalotay, E. A., Altman, E. I. (2017). Intertemporal forecasts of defaulted bond recoveries and portfolio losses. *Review of Finance*, 21(1):433–463.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., Mullainathan, S. (2017). Human decisions and machine predictions. *Quarterly Journal of Economics*, 133(1):237–293.

Koopman, S. J., Lucas, A., and Schwaab, B. (2011). Modeling frailty-correlated defaults using many macroeconomic covariates. *Journal of Econometrics*, 162(2):312–325.

Kramer, M. (1991). Nonlinear principal component analysis using autoassociative neural networks. *AIChE Journal*, 37(2):233–243.

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

Leow, M., Mues, C., Thomas, L. C. (2014). The economy and loss given default: Evidence from two UK retail lending data sets. *Journal of the Operational Research Society*, 65(3):363–375.

Loskiewicz-Buczak, A., Uhrig, R. E. (1994). Decision fusion by fuzzy set operations. In Fuzzy Systems, 1994. IEEE World Congress on Computational Intelligence., *Proceedings of the Third IEEE Conference on*, S. 1412–1417. IEEE.

Loterman, G., Brown, I., Martens, D., Mues, C., and Baesens, B. (2012). Benchmarking regression algorithms for loss given default modeling. *International Journal of Forecasting*, 28(1):161–170.

Ludvigson, S., Ng, S. (2009). Macro factors in bond risk premia. *Review of Financial Studies*, 22(12):5027–5067.

Manela, A., Moreira, A. (2017). News implied volatility and disaster concerns. *Journal of Financial Economics*, 123(1):137–162.

Meinshausen, N., Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473.

Moody's GlobalCreditResearch (2011). Still no silver bullets. *Moody's Investor Service*.

Mora, N. (2015). Creditor recovery: The macroeconomic dependence of industry equilibrium. *Journal of Financial Stability*, 18:172–186.

Nazemi, A., Pour, F. F., Heidenreich, K., Fabozzi, F. J. (2017). Fuzzy decision fusion approach for loss-given-default modeling. *European Journal of Operational Research*, 2(262):780–791.

Nazemi, A., Heidenreich, K. (2017). Artificial intelligence techniques for credit risk management. In *Intelligent Computational Systems: A Multi-Disciplinary Perspective.* Bentham Science Publishers, 268-293.

Nazemi, A., Heidenreich, K., Fabozzi, F. J. (2018a). Improving corporate bond recovery rate prediction using multi-factor support vector regressions. *European Journal of Operational Research*, 2(271):664–675.

Nazemi, A., Fabozzi, F. J. (2018). Macroeconomic variable selection for creditor recovery rates. *Journal of Banking & Finance*, 89:14–25.

Nazemi, A., Heidenreich, K., Fabozzi, F. J. (2018b). Intertemporal defaulted bond recovery prediction via machine learning. Working Paper, EDHEC Business School.

Nazemi, A., Heidenreich, K. (2018). High-dimensional analysis of macroeconomic variables' impact on recovery rates and an application to macro stress testing. Working Paper.

Nozaki, K., Ishibuchi, H., Tanaka, H. (1996). Adaptive fuzzy rule-based classification systems. *IEEE Transactions on Fuzzy Systems*, 4(3):238–250.

Park, Y. W., Bang, D. W. (2014). Loss given default of residential mortgages in a low ltv regime: Role of foreclosure auction process and housing market cycles. *Journal of Banking & Finance*, 39:192–210.

Qi, M. and Zhao, X. (2011). Comparison of modeling methods for loss given default. *Journal of Banking & Finance*, 35(11):2842–2855.

Renault, O., Scaillet, O. (2004). On the way to recovery: A nonparametric bias free estimation of recovery rate densities. *Journal of Banking & Finance*, 28(12):2915–2931.

Rösch, D. and Scheule, H. (2014). Forecasting probabilities of default and loss rates given default in the presence of selection. *Journal of the Operational Research Society*, 65(3):393–407.

Schölkopf, B., Smola,A., and Müller, K.-R. (1997). Kernel principal component analysis. In *International Conference on Artificial Neural Networks*, pages 583–588. Springer.

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

Sjöstrand, K., Clemmensen, L., Larsen, R., and Ersbøll, B. (2012). Spasm: A matlab toolbox for sparse statistical modeling. *Journal of Statistical Software.*

Storn, R., Price, K. (1997). Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization*, 11(4):341–359.

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., Zeileis, A. (2008). Conditional variable importance for random forests. *BMC bioinformatics*, 9(1):307.

Suykens, J. A. and Vandewalle, J. (1999). Least squares support vector machine classifiers. *Neural Processing Letters*, 9(3):293–300.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, S. 267–288.

Tobback, E., Martens, D., Van Gestel, T., and Baesens, B. (2014). Forecasting loss given default models: Impact of account characteristics and the macroeconomic state. *Journal of the Operational Research Society*, 65(3):376–392.

Van den Burg, G. J., Groenen, P. J., Alfons, A. (2017). Sparsestep: Approximating the counting norm for sparse regularization. *arXiv preprint arXiv:*1701.06967.

Van der Maaten, L., Postma, E., and van den Herik, H. (2007). Matlab toolbox for dimensionality reduction. *MICC, Maastricht University.*

Vapnik, V., Golowich, S. E., and Smola, A. (1997). Support vector method for function approximation, regression estimation, and signal processing. *Advances in Neural Information Processing Systems*, 281–287.

Virolainen, K. (2004). Macro stress testing with a macroeconomic credit risk model for finland. Bank of Finland discussion paper, (18).

Wang, Q. (2012) Kernel principal component analysis and its applications in face recognition and active shape models. *arXiv preprint arXiv:1207.3538.*

Yang, B. H., Tkachenko, M. (2012). Modeling exposure at default and loss given default: empirical approaches and technical implementation. *Journal of Credit Risk*, 8(2):81.

Yao, X., Crook, J., Andreeva, G. (2015). Support vector regression for loss given default modelling. *European Journal of Operational Research*, 240(2):528–538.

Yao, X., Crook, J., Andreeva, G. (2017). Enhancing two-stage modelling methodology for loss given default with *European Journal of Operational Research*, 263(2):679–689.

Zhang, C.-H. et al. (2010). Nearly unbiased variable selection under minimax concave penalty. *Annals of statistics*, 38(2):894–942.

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

Zou, H., Hastie, T., and Tibshirani, R. (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15(2):265–286.

Appendix A:

Intertemporal Defaulted Bond Recoveries Prediction via Machine Learning

Groups of independent variables part I

Seniority	
Senior unsecured	Senior secured
Senior subordinated	Subordinated
Industry	
Utility	Financials
Communication	Consumer-cylical
Industrial	IndustryDistress1
IndustryDistress2	
Bond Characteristics	
Zero Coupon	Variable Coupon
Step-up	Convertible
Insured	Retail Note
Corporate medium-term note	
News	
NewsVIX	Government
Intermediation	Natural Disaster
Securities	War
Other	
Financial Conditions	
Loans and Leases in Bank Credit, All Commercial Banks	Real Estate Loans, All Commercial Banks
Federal Debt: Total Public Debt	Total Consumer Credit Owned and Securitized, Outstanding
Excess Reserves of Depository Institutions	Commercial and Industrial Loans, All Commercial Banks
Total Borrowings of Depository Institutions from the Federal Reserve	Bank Credit of All Commercial Banks
Household Debt Service Payments as a Percent of Disposable Personal Income	Household Financial Obligations as a percent of Disposable Personal Income
Loans and Leases in Bank Credit, All Commercial Banks	Nonperforming Total Loans (past due 90+ days plus nonaccrual) to Total Loans
Nonperforming Loans to Total Loans (avg assets betw. USD 100M and 300M)	Net Loan Losses to Average Total Loans for all U.S. Banks
Total Net Loan Charge-offs to Total Loans for Banks	Return on Average Equity for all U.S. Banks
Loan Loss Reserve to Total Loans for all U.S. Banks	Nonperforming Commercial Loans (past due 90+ days plus nonaccrual) to Commercial Loans
Monetary Measures	
M2 Money Stock	Consumer Price Index for All Urban Consumers: All Items Less Food
University of Michigan Inflation Expectation	Consumer Price Index for All Urban Consumers: Energy
Personal Saving	Personal Saving Rate
Gross Saving	Gross Domestic Product: Implicit Price Deflator
Consumer Price Index for All Urban Consumers: Apparel	Consumer Price Index for All Urban Consumers: All Items
Consumer Price Index for All Urban Consumers: Medical Care	Consumer Price Index for All Urban Consumers: Transportation
Consumer Price Index for All Urban Consumers: All items less shelter	Consumer Price Index for All Urban Consumers: All items less medical care
Consumer Price Index for All Urban Consumers: Durables	Consumer Price Index for All Urban Consumers: Services
Consumer Price Index for All Urban Consumers: Commodities	Board of Governors Monetary Base, Adjusted for Changes in Reserve Requirements
M1 Money Stock	M3 for the United States
All-Transactions House Price Index for the United States	
Corporate Measures	
Corporate Profits After Tax (without IVA and CCAdj)	Corporate Profits After Tax with Inventory Valuation and Capital Consumption Adjustments
Corporate Profits after tax with IVA and CCAdj: Net Dividends	Corporate Net Cash Flow with IVA

Groups of independent variables part II

Business Cycle	
Real Gross Domestic Product	ISM Manufacturing: PMI Composite Index
Industrial Production Index	University of Michigan: Consumer Sentiment
Private Nonresidential Fixed Investment	Real Disposable Personal Income
National income	Personal Income
Manufacturing Sector: Real Output	Real Personal Consumption Expenditures
Industrial Production: Manufacturing (NAICS)	Personal Consumption Expenditures: Durable Goods
Government Consumption Expenditures & Gross Investment	Gross Private Domestic Investment
Civilian Unemployment Rate	Continued Claims (Insured Unemployment)
Average Weekly Hours of Production and Nonsupervisory Employees: Mfg	Civilian Employment
Civilian Employment-Population Ratio	Persons unemployed 15 weeks or longer, as a percent of the civilian labor force
Manufacturers' New Orders: Durable Goods	Real Final Sales of Domestic Product
Manufacturers' New Orders: Nondefense Capital Goods Excluding Aircraft	Total Business: Inventories to Sales Ratio
Capacity Utilization: Manufacturing	Change in Private Inventories
Capacity Utilization: Total Industry	Total Business Inventories
Light Weight Vehicle Sales: Autos & Light Trucks	Housing Starts: Total: New Privately Owned Housing Units Started
Housing Starts: Total: New Privately Owned Housing Units Started	New One Family Houses Sold: United States
New Private Housing Units Authorized by Building Permits	Final Sales to Domestic Purchasers
Value of Manufacturers' New Orders for Consumer Goods Industries	Value of Manufacturers' Unfilled Orders for Durable Goods Industries
Avg Weekly Overtime Hours of Production and Nonsupervisory Employees: Mfg	Avg Hourly Earnings of Production and Nonsupervisory Employees: Goods-Producing
Avg Hourly Earnings of Production and Nonsupervisory Employees: Construction	Avg Hourly Earnings of Production and Nonsupervisory Employees: Manufacturing
Real Manufacturing and Trade Industries SalesAdjusted	Personal consumption expenditures: Durable goods (implicit price deflator)
Personal consumption expenditures: Nondurable goods (implicit price deflator)	Personal consumption expenditures (implicit price deflator)
Housing Starts in Midwest Census Region	Housing Starts in Northeast Census Region
Housing Starts in South Census Region	Housing Starts in West Census Region
Initial Unemployment Claims	Real Manufacturing and Trade Inventories
Industrial Production: Business Equipment	Industrial Production: Consumer Goods
Industrial Production: Durable Consumer Goods	Industrial Production: Durable Materials
Industrial Production: Final Products (Market Group)	Industrial Production: Fuels
Industrial Production: Manufacturing (SIC)	Industrial Production: Materials
Industrial Production: Nondurable Consumer Goods	Industrial Production: Nondurable Materials
Manufacturers: Inventories to Sales Ratio	Nonfarm Private Construction Payroll Employment
Nonfarm Private Financial Activities Payroll Employment	Nonfarm Private Goods - Producing Payroll Employment
Nonfarm Private Manufacturing Payroll Employment	Nonfarm Private Service - Providing Payroll Employment
Total Nonfarm Private Payroll Employment	Nonfarm Private Trade, Transportation, and Utilities Payroll Employment
New Private Housing Units Authorized by Building Permits in the Midwest	New Private Housing Units Authorized by Building Permits in the Northeast
New Private Housing Units Authorized by Building Permits in the South	New Private Housing Units Authorized by Building Permits in the West
Number of Civilians Unemployed for 5 to 14 Weeks	Number of Civilians Unemployed for 15 Weeks and Over
Number of Civilians Unemployed for 15 to 26 Weeks	Number of Civilians Unemployed for 27 Weeks and Over
Number of Civilians Unemployed for Less Than 5 Weeks	Average (Mean) Duration of Unemployment
Consumer Opinion Surveys: Confidence Indicators: OECD Indicator for the US	Growth rate of nominal GDP
Growth rate of Nominal Dispoable Income	

Groups of independent variables part III

Stock Market	
S&P 500 Index return	S&P 500 Volatility 1m
CBOE DJIA Volatility Index	NASDAQ 100 Index return
CBOE NASDAQ 100 Volatility Index	Russell 2000 Price Index return
Russell 2000 Vol 1m	Wilshire US Small-Cap Price Index return
Wilshire Small Cap Vol	
International Competitiveness	
Real Trade Weighted U.S. Dollar Index: Broad	Trade Weighted U.S. Dollar Index: Major Currencies
Total Current Account Balance for the United States	Real Exports of Goods & Services
Balance on Merchandise Trade	Real imports of goods and services
Canada / U.S. Foreign Exchange Rate, Canadian Dollars to One U.S. Dollar	Japan / U.S. Foreign Exchange Rate, Japanese Yen to One U.S. Dollar
Switzerland / U.S. Foreign Exchange Rate, Swiss Francs to One U.S. Dollar	U.S. / U.K. Foreign Exchange Rate, U.S. Dollars to One British Pound
Real Broad Effective Exchange Rate for United States	
Micro-level	
Manufacturing Sector: Unit Labor Cost	Nonfarm Business Sector: Unit Labor Cost
Compensation of employees: Wages and salaries	Compensation of employees: Mfg: Nondurables: Food, beverage and tobacco
Employment Cost Index: Total comp in Management, professional, and related	Manufacturing Durable Goods Sector: Compensation
Employment Cost Index: Benefits: Private Industry Workers	Employment Cost Index: Total comp for civilian workers in all industries and occupations
Employment Cost Index: Wages & Salaries: Private Industry Workers	1-Month AA Nonfinancial Commercial Paper Rate
10-Year Treasury Constant Maturity Rate	3-Month AA Nonfinancial Commercial Paper Rate
TermStructure	Effective Federal Funds Rate
Moody's Seasoned Baa Corporate Yield Relative to Yield on 10-Year Treasury	Moody's Seasoned Aaa Corporate Bond Yield
30-Year Conventional Mortgage Rate	Moody's Seasoned Baa Corporate Bond Yield
Bank Prime Loan Rate	Producer Price Index for All Commodities
Producer Price Index by Commodity Industrial Commodities	Producer Price Index by Commodity Intermediate Energy Goods
Producer Price Index by Commodity for Crude Energy Materials	Producer Price Index by Commodity for Finished Consumer Goods
Producer Price Index by Commodity Intermediate Materials	6-Month Treasury Bill: Secondary Market Rate
1-Year Treasury Constant Maturity Rate	5-Year Treasury Constant Maturity Rate
3-Month Treasury Bill: Secondary Market Rate	3-month Treasury Constant Maturity Rate
Producer Price Index by Commodity for Final Demand: Finished Goods	Moody's Seasoned Aaa Corporate Bond Minus Federal Funds Rate
Moody's Seasoned Baa Corporate Bond Minus Federal Funds Rate	3-Month Commercial Paper Minus Federal Funds Rate
Moody's Seasoned Aaa Bbb Spread	Size of High Yield Market in U.S. Dollars
High Yield Default Rate, Trailing 12-month	Bond defaults within the industry (in percent)