ARTICLE IN PRESS
JID: EOR [m5G;June 21, 2018;1:35]

European Journal of Operational Research 000 (2018) 1–12

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Innovative Applications of O.R.

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

Abdolreza Nazemi[a], Konstantin Heidenreich[a], Frank J. Fabozzi[b,*]

[a] School of Economics and Business Engineering, Karlsruhe Institute of Technology, Germany
[b] EDHEC Business School, Nice, France

ARTICLE INFO

Article history:
Received 10 October 2017
Accepted 10 May 2018
Available online xxx

Keywords:
Recovery rate
Least-squares support vector regression methods
Sparse principal component analysis
Kernel principal component analysis
Gradient boosting

ABSTRACT

In the multi-factor framework described in this paper, we use instrument-specific characteristics, several macroeconomic variables, and industry-specific characteristics as our explanatory variables for predicting recovery rates for corporate bonds. By including the principal components derived from a large number of macroeconomic variables, all three least-squares support vector regression methods, as well as the ordinary linear regression, exhibit higher out-of-sample predictive accuracy than the models that included only the few macroeconomic variables suggested in the literature. We compare the prediction accuracies of all techniques by incorporating sparse principal components, nonlinear principal components from an auto-associative neural network, and kernel principal components. Our results show that sparse principal components generate more interpretable and accurate estimations compared to the other principal component techniques. Moreover, we apply gradient boosting to generate a ranking of the 104 macroeconomic variables, from best to worst, based on their prediction power in recovery rate estimation. The three categories with the most informative macroeconomic predictors are micro-level factors, business cycle variables, and stock market indicators.

1. Introduction

Enhanced 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. However, as a consequence of the Basel II accord, loss given default has become a much more critical measure for banks and other financial institutions.

According to Loterman, Brown, Martens, Mues, and Baesens (2012) within the required framework of Basel II, loss given default has a linear impact on the required minimum capital. Better prediction models allow for the calculation of more realistic capital requirements, as well as providing more accurate valuations for 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, Brady, Resti, and Sironi (2005) conclude that recovery rates of corporate bonds are related to default rates, seniority and collateral levels. Acharya, Bharath, and Srinivasan (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. Jankowitsch, Nagler, and Subrahmanyam

* Corresponding author.
E-mail addresses: abdolreza.nazemi@kit.edu (A. Nazemi), konstantin.heidenreich@gmx.de (K. Heidenreich), fabozzi321@aol.com, frank.fabozzi@edhec.edu (F.J. Fabozzi).

https://doi.org/10.1016/j.ejor.2018.05.024
0377-2217/

Please cite this article as: A. Nazemi et al., Improving corporate bond recovery rate prediction using multi-factor support vector regressions, European Journal of Operational Research (2018), https://doi.org/10.1016/j.ejor.2018.05.024

(2014) conclude that the magnitude and variability of defaults during the global financial crisis have increased the importance of calculating accurate forecasts of loss given default. In a low default world, inaccurate forecasts might have been less risky and might not have been detected for the simple reason that less default events occurred.

In the paper, we use several factors such as macroeconomic variables, instrument-specific characteristics, and industry-specific 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. Presenting a comprehensive comparison of 24 techniques for the prediction of recovery rates of diverse instruments such as corporate loans, mortgage loans and personal loans, Loterman et al. (2012) show a clear trend that non-linear techniques such as support vector machines and artificial neural networks have more predictive power than traditional linear models. Moreover, they argue that two-stage models – a combination of non-linear and linear models – have similar predictive power like non-linear models with the advantage of having a comprehensible linear model component. The other study is by Yao, Crook, and Andreeva (2015) who apply support vector methods to predict recovery rates for corporate bonds. In their study using three variations of a least-squares support vector regression (LS-SVR), they report significant outperformance compared to traditional modeling techniques such as linear regression or fractional response regression. Moreover, they also report that LS-SVR shows outperformance at a segmented level by bond seniority versus traditional approaches.

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 need not be calculated explicitly. After mapping to high-dimensional linear space, SVR can provide more accurate predictions. 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 applying multi-factor SVR. We compare the predictive performance using different data reduction techniques for 104 macroeconomic variables such as principal component analysis (PCA), sparse PCA, nonlinear PCA, kernel PCA 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, Martens, Van Gestel, and Baesens (2014) highlight the importance of macroeconomic independent variables when modeling recovery rates for corporate loans and home equity loans. By adding 11 macroeconomic indicators, they are able to improve the performance of their models which include SVR, a regression tree, a linear model and a two-stage model combining a linear approach with SVR, significantly. Both Duffie, Eckner, Horel, and Saita (2009) and Koopman, Lucas, and Schwaab (2011) show that macroeconomic influences matter a lot for calculating the probability of default. In particular, they demonstrate the impact of a latent, dynamic frailty factor. To make sure the frailty factor captures only really unobservable effects, Koopman et al. (2011) include 10 principal components derived from more than 100 macroeconomic variables in their model.

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 variance in the dataset 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 overview of our multi-factor framework and the variables selected are provided in the next section, Section 2. We also present the data reduction techniques we apply to the 104 macroeconomic variables. In Section 3 we describe how and why we have chosen our modeling techniques which are SVR, regression tree and linear regression. An exploratory analysis of our dataset consisting of 775 corporate bonds with default events between 2002 and 2012 is presented in Section 4. The out-of-sample performance of our cross-validated models is discussed in the middle part of Section 4, where we show that LS-SVR models have a higher predictive capacity. In particular, our models' performance is increased by adding the principal components of 104 macroeconomic variables. In the last part of Section 4, we discuss the macroeconomic variables' ranking by gradient boosting and the predictive accuracies when including the highest ranked macroeconomic variables. Section 5 concludes our paper.

2. Multi-factor framework

In this section we describe our multi-factor framework, its inputs, and its extension by adding principal components of macroeconomic variables.

2.1. Selection of factors for modeling

We define the recovery rate r_{ij} for bond i of firm j in our framework as follows:

$$r_{ij} = \alpha + \beta_c X_{ci} + \beta_m X_{mi} + \beta_{ind} X_{indj} + \epsilon_{ij}, \quad (1)$$

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;