European Journal of Operational Research 297 (2022) 1162–1177

Contents lists available at ScienceDirect

European Journal of Operational Research

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

Interfaces with Other Disciplines

Intertemporal defaulted bond recoveries prediction via machine learning[☆]

Abdolreza Nazemi[a], Friedrich Baumann[b], Frank J. Fabozzi[c,*]

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

ARTICLE INFO

Article history:
Received 12 March 2020
Accepted 22 June 2021
Available online 29 June 2021

JEL classification:
G33
G17
C14
C52

Keywords:
Finance
Risk management
Recovery rates
Machine learning
News-based analysis
Power expectation propagation

ABSTRACT

The recovery rate on defaulted corporate bonds has a time-varying distribution, a topic that has received limited attention in the literature. We apply machine learning approaches for intertemporal analysis of U.S. corporate bonds' recovery rates. We show that machine learning techniques significantly outperform traditional approaches not only out-of-sample as documented in the literature but also in various out-of-time prediction setups. The newly applied sparse power expectation propagation approach provides the most compelling out-of-time prediction results. Motivated by the association of systematic factors with the time-varying characteristic of recovery rates, we study the effect of text-based news measures to account for bond investors' expectations about the future which translate into market-based recovery rates. Especially during recessions, government-related news are associated with higher recovery rates. Although machine learning is a data-driven approach rather than considering economic intuition for ranking a group of predictors, the most informative groups of predictors for recovery rate prediction are nevertheless economically meaningful.

1. Introduction

The determinants of recovery rates play an important role in the valuation of default risk insurance. Moreover, the advanced internal ratings-based approach under the Basel Accord II and III allows financial institutions to use their own estimates for credit risk parameters. Consequently, accurate and reliable estimates for recovery rates are needed. Although studies by practitioners and academics have investigated recovery rate determinants of defaulted corporate bonds and loans, as well as alternative prediction methods for estimating recovery rates, these studies seldomly fully account for the time variation of recovery rates.

Recent studies have examined out-of-sample or in-sample settings to analyze the determinants of recovery rates.[1] According to Kalotay & Altman (2017), the applicability of conventional out-of-sample estimation to the field of recovery rate prediction is questionable. In particular, the *k*-fold cross-validation method, the commonly used performance measure for evaluating the predictive accuracy for the recovery rate, has shortcomings. For the *k*-fold cross validation method the dataset is randomly divided into *k* subsamples. Each subsample is used for out-of-sample prediction once, while the remaining *k*-1 subsamples are used for training. The performance measurement is the average of the predictions for the *k*-th subsample.

Even though conventional out-of-sample estimation has been established as the standard procedure in academia, only out-of-time prediction is feasible in real-world applications of corporate debt recovery rate prediction. While it is acknowledged that out-of-sample estimation makes a distinction between training and testing data for recovery rate prediction, it suffers from two main shortcomings. First, as the dataset is randomly partitioned for out-of-sample estimation, it is virtually inevitable that defaults of bonds used for training the model have occurred after defaults of bonds used for testing the model. Consequently, out-of-sample prediction assumes the data-generating process to be time-invariant, leading to a look-ahead bias in the predictions. The sec-

[☆] We gratefully thank Dr. Konstantin Heidenreich whose support and inputs were very valuable for completing this study.

[*] Corresponding author.

E-mail addresses: abdolreza.nazemi@kit.edu (A. Nazemi), friedrich.w.baumann@gmail.com (F. Baumann), frank.fabozzi@edhec.edu (F.J. Fabozzi).

[1] See, for example, Qi & Zhao (2011); Yao, Crook, & Andreeva (2015) and Nazemi & Fabozzi (2018).

https://doi.org/10.1016/j.ejor.2021.06.047
0377-2217/© 2021 Published by Elsevier B.V.

ond shortcoming is that out-of-sample prediction implicitly makes the questionable assumption that recovery rates of two defaulted bonds issued by the same company are independent of each other. For instance, when two bonds from the same issuer have defaulted at the same time, only for out-of-time estimation it is ensured that these two bonds are either both in the training set or both in the test set, such that the recoveries in the test set are independent from the training set.[2]

In this paper, we address these shortcomings by comparing a wide range of statistics and machine learning methods – 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. We predict out-of-time by ensuring that only sample points observed before the default event were used during the training process (see Section 5.3 for more details).

The literature on recovery rates has identified both cross-sectional factors that are time-invariant and systematic factors with a time-varying dependency as determinants of recovery rates. In order to create better models of recovery rates, new insights into the drivers of the time variation of recovery rates are needed.[3] We extend the sparse existing research devoted to modeling recovery rates' time variation. Therefore, we include text-based measures extracted from front-page articles published in *The Wall Street Journal* as independent variables. By including these text-based measures, we consider aggregate uncertainty about future economic conditions prevalent at the time of default. As bond prices shortly after default represent expected recovery, these are driven by investors' expectations about future cash flows. These, in turn, are subject to uncertainty, hence the time-variant distribution of recovery rates is ultimately connected to economic uncertainty. Furthermore, we compare selection techniques such as stability selection, MC+ algorithm, and SparseStep algorithm for selecting a subset of macroeconomic variables from a large set of macroeconomic measures in order to identify those which are most closely related to the recovery rate. This study is the first to compare these econometric and machine learning methods in empirical finance.

Our primary contribution to the recovery rate literature is threefold. First, in addition to presenting a machine learning framework for out-of-sample recovery rate prediction, we evaluate the intertemporal prediction performance of a wide range of parametric and non-parametric techniques across various out-of-time prediction setups. Surprisingly, these have attracted less attention in the literature than out-of-sample prediction techniques. 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, this study is the first to apply sparse power expectation propagation for modeling the recovery rate. The best out-of-time prediction accuracy is achieved using a sparse power expectation propagation approach, outperforming support vector regression-based approaches by Yao et al. (2015) and Nazemi, Heidenreich, & Fabozzi (2018) published in this journal. Third, our study includes news-based measures that have been extracted and categorized with machine learning techniques as an alternative group of independent variables to account for the time variation in recovery rate estimation. By incorporating news-based variables, we show that these variables are significant drivers of recovery rates.

Besides these main contributions, this study is the first to benchmark several selection techniques from the machine learning and econometrics literature for out-of-sample identification and selection of the most informative macroeconomic variables for recovery rate prediction from high-dimensional data. To the best of our knowledge, this study is the first to investigate a ranking of the groups of all independent variables including bond characteristics, seniority, news-based, industry, and seven groups of macroeconomic variables. This ranking, from most informative to least informative, is based on the groups' permutation importance for predicting recovery rates of U.S. corporate bonds with the random forest method. It provides the interesting insight that some of the most informative variables for recovery rate prediction have attracted less attention from previous research than their importance suggests.

We organize the remainder of the paper as follows. A review of the literature is presented in Section 2. In Section 3 we describe the modeling techniques and selection algorithms we applied. We describe the data we used in Section 4 and present our empirical results in Section 5. Our conclusions are provided in Section 6.

2. Literature review

Altman & 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 after controlling for seniority, the original rating of a defaulted bond has no impact on the recovery rate. Altman, Brady, Resti, & Sironi (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 a weak correlation with the average recovery rate. Acharya, Bharath, & Srinivasan (2007) document that creditors recover less if the industry of the defaulted firm is in distress. In particular, they show using a dataset for the years 1982 to 1999 that defaulted corporate bonds in distressed industries exhibit 10% to 15% lower average recovery rates.

Altman & 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. In an in-sample study, Jankowitsch, Nagler, & Subrahmanyam (2014) examine the recovery rates of defaulted bonds while paying special attention to the trading microstructure around various types of default events. They find in an in-sample analysis that (1) trading volume in the 30 days after the default is high while trading activity decreases after this time period and (2) bond characteristics (e.g., coupon and covenants) and CDS availability have a significant impact on market-based recovery rates.

Jansen, Das, & Fabozzi (2018) and Schläfer & Uhrig-Homburg (2014) use the term structure model for the recovery rate of credit default swaps. Calabrese & Zenga (2010) suggest a beta regression model for the estimation of bank loan recovery rates. Hartmann-Wendels, Miller, & Töws (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. Yao, Crook, & Andreeva (2017) suggest incorporating a two-stage modeling framework to predict recovery rates of credit cards. Krüger & Rösch (2017) study the downturn loss-given-default employing the quantile regression technique for both in-sample and out-of-sample estimation. Hurlin, Leymarie, & Patin (2018) apply six mod-

[2] In this respect, recovery rates of corporate bonds and loans are different to recovery rates of consumer credit as these have a less time-varying distribution and the interdependence of multiple defaults is a minor aspect. Betz, Kellner, & Rsch (2021) investigate how default resolution times impact final loss rates of loans.

[3] See, for example, Doshi, Elkamhi, & Ornthanalai, 2018; Kalotay & Altman, 2017.

els for modeling LGD of almost 10,000 defaulted Brazilian credit and leasing contracts. Cheng & 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 Acharya et al. (2007); Jankowitsch et al. (2014); Qi & Zhao (2011); Varma & Cantor (2005); Yao et al. (2015) use only a few macroeconomic variables. Nazemi et al. (2018) report that models for estimating recovery rates significantly outperform by adding principal components derived from 104 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 dataset applied 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. In our study, we expand their work by comparing the empirical performance of several selection techniques.

The literature documents time variation of corporate bond recovery rates and demonstrates that the factors which explain recovery rates depend on both cross-sectional features, e.g. bond-specific characteristics, as well as on systematic features, e.g. macroeconomic time series data.[4] Doshi et al. (2018) note that there is a general agreement among academics that corporate debt recovery rates are time-varying, but the empirical work on this characteristic is limited. Chen (2010), highlighting the dependence of recovery rates' time variation on the effects of systematic variables such as the economic cycle, states that the average values of recovery rates during recessions (1982, 1990, 2001, and 2008) are smaller than during economic upswings. Bruche & Gonzalez-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 an unobserved Markov chain to account for time variation in the probability of default and the recovery rate. Nazemi & Baumann (2021) demonstrate the relation of corporate bond recovery rates to time-varying stock- and bond-risk factors. Gambetti, Gauthier, & Vrins (2019) employ the concept of economic uncertainty in recovery rate prediction and find it to explain a larger fraction of the systematic variation of recovery rates than other time-variant proxies for the business cycle. The authors explain the relationship between uncertainty and recovery by arguing that recovery, measured by the bond price shortly after default, represents investors' expectations about future cash flow, which in turn depends on economic conditions and is therefore subject to uncertainty. In their study, they employ different uncertainty measures and find that increasing economic uncertainty is significantly connected to decreasing recovery rates even when controlling for the business cycle. One of their explanatory variables, an economic policy uncertainty measure (EPU) constructed by Baker, Bloom, & Davis (2016), considers news-based uncertainty that captures the monthly number of newspaper articles containing specific expressions related to economic policy uncertainty. Baker et al. (2016) create this measure through human audit, analyzing over 12,000 newspaper articles by hand and manually selecting policy-related expressions.

We incorporate five text-based news measures created by Manela & Moreira (2017), who emphasize the possibility of explaining asset price fluctuations with time-varying levels of uncertainty that are contained within news articles. They create their measures by extracting information from front-page articles published in *The Wall Street Journal* using machine learning techniques. Their approach is based on the co-movement of word frequencies with option-implied volatility. Using *WordNet* and *WordNet::Similarity* for classification, they construct five interpretable word categories. Whereas Baker et al. (2016) exclusively consider news which specifically contain the words "uncertainty" or "uncertain" and therefore create a measure that, by intuition, has a negative connotation, the methodology of Manela & Moreira (2017) doesn't capture any predefined sentiment. Because the news categories which they identify are not unique to uncertainty from economic policy concerns, their measures account for more diverse sources of time-varying uncertainty. They further argue that their approach does not depend on human interaction or judgment in the expression selection process, and is therefore more objective than the EPU measure. While the text-based news measures are able to capture uncertainty that is priced by the stock market, EPU does not.[5]

Compelled by the sparse literature on out-of-time estimation of recovery rates, Kalotay & Altman (2017) investigate the time variation of recovery rates.[6] They report that parametric methods outperform the non-parametric methods for intertemporal prediction 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 is employed in their analysis.

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 particular, 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. Nazemi, Fatemi Pour, Heidenreich, & Fabozzi (2017); Qi & Zhao (2011); Yao et al. (2015) and Nazemi et al. (2018) report that that non-parametric techniques such as regression trees and support vector regressions outperform parametric methods for predicting recovery rates of corporate bonds in an out-of-sample study. Table 1 provides an overview of the prevalence of recovery rate model validation techniques in recent studies for U.S. corporate bonds. The k-fold cross-validation is the most frequently used method for out-of-sample evaluation and has been established as standard procedure in recovery rate estimation. The k-fold cross-validation has two mains drawbacks described in Section 1. None of these studies make comparisons between machine learning techniques and statistical models with attention to intertemporal forecasting performance, and with the exception of Kalotay & Altman (2017) who compared statistical models with just regression trees. Our principal contribution relative to the literature applies yet unrecognized machine learning techniques for intertemporal analysis of U.S. corporate bonds' recovery rates.

[4] Several studies emphasize the time-varying characteristic of corporate debt recovery rates. Varma & Cantor (2005) mention considerable time-variation for recovery rates of defaulted bonds and loans. Acharya et al. (2007) describe the time-series behavior of recovery rates of defaulted loans and bonds over the period 1982–1999. Jankowitsch et al. (2014) mention that recovery rates exhibit substantial variation over time. Mora (2015) evaluates recovery rates of bonds, loans and preferred stock within the time-dependent cyclicality of economic conditions.

[5] Cortes & Weidenmier (2019) also find no significant association of EPU with stock volatility.

[6] Besides this paper, Doshi et al. (2018) estimate a time-varying recovery rate of credit default swaps.

Table 1
Overview of recovery rate models for U.S. corporate bonds in literature.

Author(s)	Data	Model(s)	In-sample	Out-of-sample	Out-of-time
Frye (2000)	Corporate bonds (1983-1997)	Conditional model	yes	no	no
Altman et al. (2005)	Corporate bonds (1982–2002)	Univariate and multivariate, logistic, logarithmic and linear regression	yes	no	no
Varma & Cantor (2005)	Bonds and loans from c. 1100 corporate issuers (1983–2003)	Univariate and multivariate regression	yes	no	no
Acharya et al. (2007)	1511 loans and bonds of over 300 firms (1982–1999)	Multivariate regression	yes	no	no
Bruche & Gonzalez-Aguado (2010)	2000 bonds (1974–2005)	Markov chain	yes	yes	no
Chava et al. (2011)	Corporate bonds (1980–2008)	Linear, logit and probit	yes	yes	no
Jacobs & Karagozoglu (2011)	Corporate loans and bonds (1985–2008)	Beta-link generalized linear model	yes	yes	no
Qi & Zhao (2011)	3751 defaulted bonds and loans (1985–2008)	Regression, fractional response regression, transformation regressions, tree, neural network	yes	yes	no
Altman & Kalotay (2014)	4720 debt instruments, of which 60% are bonds (1987–2011)	Parametric regressions, regression trees Bayesian conditional mixture	yes	yes	yes
Jankowitsch et al. (2014)	Corporate bonds, 1270 default events of 534 firms (2002–2010)	Multivariate regression	yes	no	no
Donovan, Frankel, & Martin (2015)	Several instruments of 347 firms (1994–2011)	Univariate and multivariate regression	yes	no	no
Mora (2015)	4422 bonds, loans and preferred stock (1970–2008)	Univariate and multivariate regression	yes	no	no
Yao et al. (2015)	1413 corporate bonds (1985–2012)	Linear regression, fractional response regression, SVRs, two-stage model	yes	yes	no
Kim & Kung (2016)	Secured corporate loans and bonds (1989–2009)	Multivariate linear regressions	yes	no	no
Kalotay & Altman (2017)	2828 non-financial corporate bonds (1987–2011)	Inverse Gaussian regressions, mixture models, regression trees	yes	yes	yes
Nazemi & Fabozzi (2018)	794 corporate bonds (2002–2012)	Linear regression, SVRs, bagging, boosting, LASSO, ridge regression	yes	yes	no
Gambetti et al. (2019)	1831 corporate bonds (1990–2013)	Beta regression, mixture models, regression trees	yes	no	no

Our study is closest to the study by Kalotay & Altman (2017) and Nazemi & Fabozzi (2018). We provide five main contributions as compared to previous studies. First, our paper investigates the ability of parametric and non-parametric models to predict recovery rates for corporate bonds in several out-of-time prediction setups, as well as out-of-sample. In contrast to Kalotay & Altman (2017), we find that machine learning techniques also outperform in intertemporal prediction. Second, we introduce the sparse power expectation propagation method to the credit risk literature and find that it yields the most compelling results for out-of-time recovery rate prediction, thereby outperforming support vector regression methods which are the most accurate methods applied in Nazemi & Fabozzi (2018); Yao et al. (2015) and Nazemi et al. (2018). Third, our paper includes several text-based variables of time-varying uncertainty constructed via machine learning techniques for explaining the recovery rates of U.S. corporate bonds. We rely on five different news categories and extend the uncertainty concept in recovery rate prediction compared to previous work. Fourth, whereas Nazemi & Fabozzi (2018) extensively investigate macroeconomic variables in recovery rate prediction and apply the LASSO for identifying and selecting the most informative macroeconomic predictors, we benchmark the performances of several advanced selection techniques for selecting macroeconomic variables from a large set of macroeconomic variables. Lastly, we investigate the permutation importance of groups of explanatory variables in order to rank the major determinants of recovery rates.

3. Corporate bond recovery rate modeling

For recovery rate modeling, we apply linear regression, inverse Gaussian regression, regression tree, random forest, semiparametric least-squares support vector regression, and power expectation propagation techniques. Our benchmark model is linear regression. In the following, we describe the power expectation propagation technique which we apply in recovery rate modeling for the first time in the literature. We provide a description of the other modeling techniques, the macroeconomic variables selection techniques utilized (LASSO, SparseStep, and MC+ algorithm), as well as a variable ranking method in Internet Appendix A. In our recovery rate models, we control for recovery rates determinants with dummy variables for industry, seniority, coupon type, and instrument type. We further use industry distress dummy variables and the news-based measures that can capture uncertainty about the future.

3.1. Power expectation propagation

According to Bui, Yan, & Turner (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. They introduce a unifying framework for sparse Gaussian process pseudo-point approximation using power expectation propagation.[7] 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 approaches zero, the power expectation propagation solution becomes the minimum of a variational free energy approach. In contrast, when alpha is equal

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

Table 2
Descriptive statistics of the recovery rates for all bonds (Panel A) and across seniority classes (Panel B). 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	2,079
Panel B: Recovery rates across seniority								
Senior Unsecured	46.25	34.51	7.50	10.00	48.00	71.00	95.41	1,715
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

to one, 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. Data

We merge several data sources such as S&P Capital IQ, Bloomberg, Federal Reserve Bank of St. Louis, and news from front-page articles of *The Wall Street Journal* for analyzing the recovery rate of U.S. corporate bonds in this study. Our initial dataset which consists of 2,080 bonds that defaulted between 2001 and 2016 is retrieved from the S&P Capital IQ database (Capital IQ). Bond data are retrieved from S&P Capital IQ. We consider market-based recovery rates based on bond pricing data 30 days after default which we retrieve from Capital IQ. The bond prices available through Capital IQ are obtained from the Intercontinental Exchange (ICE) and are, depending on availability of respective data sources, based upon evaluations of dealer quotes, live trading levels and trade execution data from the Trade Reporting and Compliance Engine (TRACE).[8] In our sample, there is only one default event observation for each bond. 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 the bonds in our study were assigned to the following industries: industrial, consumer discretionary, consumer staples, telecommunications, raw materials, utilities, energy, financial services, and information technology. We exclude one bond from our analysis because the data was corrupt.

The remaining 2,079 bonds exhibit an average recovery rate of 45.57% and a sample standard deviation of 35.04% as reported in Table 2. We combine the seniority classes junior subordinate and subordinate to one class because these two classes contain the fewest bonds. 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 of 8.15%, while senior subordinated bonds have the second lowest average recovery rate. Accordingly, senior secured bonds have the highest average recovery rate of 61.91%. 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%).

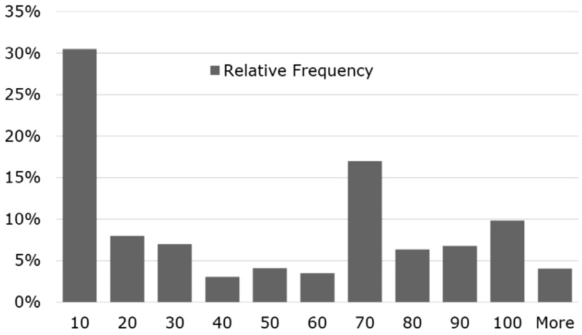

Fig. 1. Relative frequency of the recovery rates for the defaulted U.S. corporate bonds from 2001 to 2016.

The histogram of the relative frequency of the observed recovery rates in our sample exhibits two peaks and is presented in Fig. 1. The class between 0% and 10% contains around 640 defaulted bonds. There is another peak of the distribution at the class of values between 60% and 70%. However, the observed distribution does not look like a bimodal distribution.

Evidence for the importance of macroeconomic variables in credit risk management can be found in Bruche & Gonzalez-Aguado (2010); Chava, Stefanescu, & Turnbull (2011); Jankowitsch et al. (2014); Mora (2015); Nazemi et al. (2017); Varma & Cantor (2005), and Nazemi & Fabozzi (2018). We use the database from the Federal Reserve Bank of St. Louis (FRED, Federal Reserve Economic Data) complemented by aggregate default data from Fitch Ratings to retrieve 182 macroeconomic variables used in the credit risk literature such as Acharya et al. (2007); Jankowitsch et al. (2014); Mora (2015); Varma & Cantor (2005), and Nazemi & Fabozzi (2018). The macroeconomic data covers the time period from 2000 (one year before the start of the recovery rate observation period) until 2016. We list the macroeconomic variables in Internet Appendix A.

A novel data source for recovery rate estimation is news from front-page news articles of *The Wall Street Journal*. To the best of our knowledge, our study is the first study to use any kind of text-based variable constructed via machine learning techniques in recovery rate estimation. News ideally fit intertemporal prediction setups since the data needed can be easily collected immediately by keeping track of the media, whereas the most recent macroeconomic data may not be available at the desired prediction date as it is often published only with a time lag. We choose the text-based news measures that were shown to reflect investors' concerns about the future in a study by Manela & Moreira (2017). They incorporate a measure of the investors' mood that goes beyond commonly used hard data. The relationship between investors' uncertainty and implied volatility is also robust after controlling for realized stock market volatility. Their work is based on the premise that news reflect the interest of readers and that words used by the business press express the concerns of the average investor.

[8] While Khieu, Mullineaux, & Yi (2012) argue that the recovery rate based on 30-day-prices is biased, Metz, Sorensen, Keisman, & Chiu (2012) find that the bond price 30 days after default serves as a powerful predictor of the mean and variability of ultimate recovery. Moreover, the 30-day market convention represents the actual recovery for investors who sell timely after default, and it provides the advantage to be observable early after default while ultimate recovery can only be observed after the issuer's emergence from default (see, for example, Mora (2015)).

Table 3
Correlation matrix of text-based news variables and the economic policy uncertainty (EPU) measure from 2001 to 2012.

	Government	Intermediation	Securities Markets	War	Unclassified	EPU
Government	1.00					
Intermediation	0.44	1.00				
Securities Markets	0.58	0.43	1.00			
War	-0.11	-0.17	-0.07	1.00		
Unclassified	0.41	0.61	0.57	0.17	1.00	
EPU	0.38	0.19	0.40	0.27	0.68	1.00

Manela & Moreira (2017) create their text-based measures by first collecting headlines and abstracts of *The Wall Street Journal* articles, then decomposing them into one- and two-word n-grams, and eventually using out-of-sample support vector regression to regress the commonly used implied volatility indices, the CBOE Volatility Index (VIX) and the CBOE OEX Implied Volatility Index (VXO), on monthly normalized n-grams. By doing so, they yield a forward-looking news-implied volatility measure that is capable of serving as a proxy for investor uncertainty. The authors demonstrate this capability by tracing back news headlines and abstracts starting in 1889, then creating a long history of news-implied volatility and showing that their measure peaks in a number of market turmoils and economic crises. In the next step, Manela and Moreira classify words in the front-page articles to determine the different sources of uncertainty inherent in the news articles. Therefore, they apply commonly used text analytics methods *WordNet* and *WordNet::Similarity*, by which the words are classified according to semantic similarity and relatedness. This step yields five distinct measures from different news categories: government, intermediation, securities markets, war, and unclassified news. The news categories can represent various sources of uncertainty. Manela and Moreira provide possible explanations, e.g. they associate news from the war category with uncertainty about the destruction of human and physical capital, whereas they relate government news to uncertainty about policy changes. Intermediation news are connected to banking crises and bank failure, and securities markets news to stock price movements.

We further obtain the EPU measure which captures economic policy uncertainty derived from manually screened news articles provided by Baker et al. (2016) in order to control for the effects from EPU on recovery rates in our analysis. Baker et al. (2016) rely on ten of the largest U.S. newspapers and therein count the number of occurrences of three policy-related term categories: “uncertainty” or “uncertain”, “economic” or “economy”, and policy-related terms “Congress”, “deficit”, “Federal Reserve”, “legislation”, “regulation”, or “White House”. The selection of these policy terms is the result of an extensive supervised human audit of over 12,000 news articles, in which the auditors manually code EPU = 1 or EPU = 0, depending on the perceived economic uncertainty. The authors then identify policy terms that appear frequently in the articles coded as contributing to economic policy uncertainty and select the above conceptual term sets.

We merge our dataset with news-based measures that are reported by Baker et al. (2016) as well as by Manela & Moreira (2017). We use the monthly time series data as a gauge for investors’ uncertainty. In Fig. 2, we plot the EPU as well as the text-based news measures. The correlation matrix in Table 3 demonstrates that there is only limited correlation between EPU and the text-based news variables in our data, suggesting that the latter contain information that is distinct from the information conveyed via the EPU measure.

5. Empirical analysis of recovery rates’ prediction

We first examine the relation between news and the recovery rate. Second, we select macroeconomic variables for recovery rate prediction and benchmark advanced selection techniques. Third, we investigate the recovery rate estimation in out-of-sample and out-of-time (intertemporal) settings. Finally, we rank the groups of explanatory variables by their permutation importance for recovery rate prediction.

5.1. Analyzing the news’ impact on recovery rates

In Table 4 we present an overview of the linear regression specifications based on the entire dataset of 2,079 corporate bonds. The recovery rate of the defaulted U.S. corporate bond is the dependent variable. In model (1), we consider a basic specification including seniority dummies, industry variables, and bond characteristics as independent variables. In model (2), we add the EPU measure from Baker et al. (2016) to the basic specification. In contrast, in model (3) we consider the five news-related measures from Manela & Moreira (2017). In model (4), we use the seven macroeconomic variables selected by stability selection in addition to seniority, industry, and bond variables. Finally, we combine the independent variables from models (3) and (4) in model (5).

Combining news-related and macroeconomic variables in model (5) generates a further improvement of in-sample fit to an adjusted R-squared of 48.26%. We show the significance of three out of five text-based measures of news even when controlling for the effects of macroeconomic variables in model (5). Similar to Acharya et al. (2007), industry distress variables have a significant negative impact on recovery rates. In line with the results from Varma & Cantor (2005), bonds with a higher payment rank in the seniority structure recover on average significantly more than bonds with a lower rank. Finally, we confirm the significance of bond characteristics reported by Jankowitsch et al. (2014).

The intuition behind applying news-based variables to recovery rate prediction is that news-based variables, which were shown to reflect investors’ time-varying concerns in previous asset pricing research, are likely to also drive the prices of defaulted debt securities. Given that the 30-day bond prices as the recovery rate represent expected future cash flows, we anticipate it to depend on the news that proxy investors’ expectations about the future.

Following Gambetti et al. (2019), we add the EPU measure to the basic specification in model (2). Gambetti et al. (2019) argue that the bond price after default as a representation of recovery reflects expectations about future cash flows, and as such is influenced by uncertainty. Furthermore, they emphasize a cause-effect between uncertainty and economic downturns, an economically coherent notion that surfaces in their observations of a significant negative impact of uncertainty on recovery. Similar to their study, we observe a significant and negative impact of the EPU on recovery rates. However, we find only a very marginal improvement of recovery estimation accuracy. We then replace the EPU with text-based variables as measures for investors’ expectations about the future in model (3). By doing so, we include a more diverse universe of uncertainty for recovery rate estimation that relies not only on uncertainty related to economic policy. The estimation accuracy improves by about 3.5 percentage points, showing that news-based variables outperform the EPU in recovery rate predic-

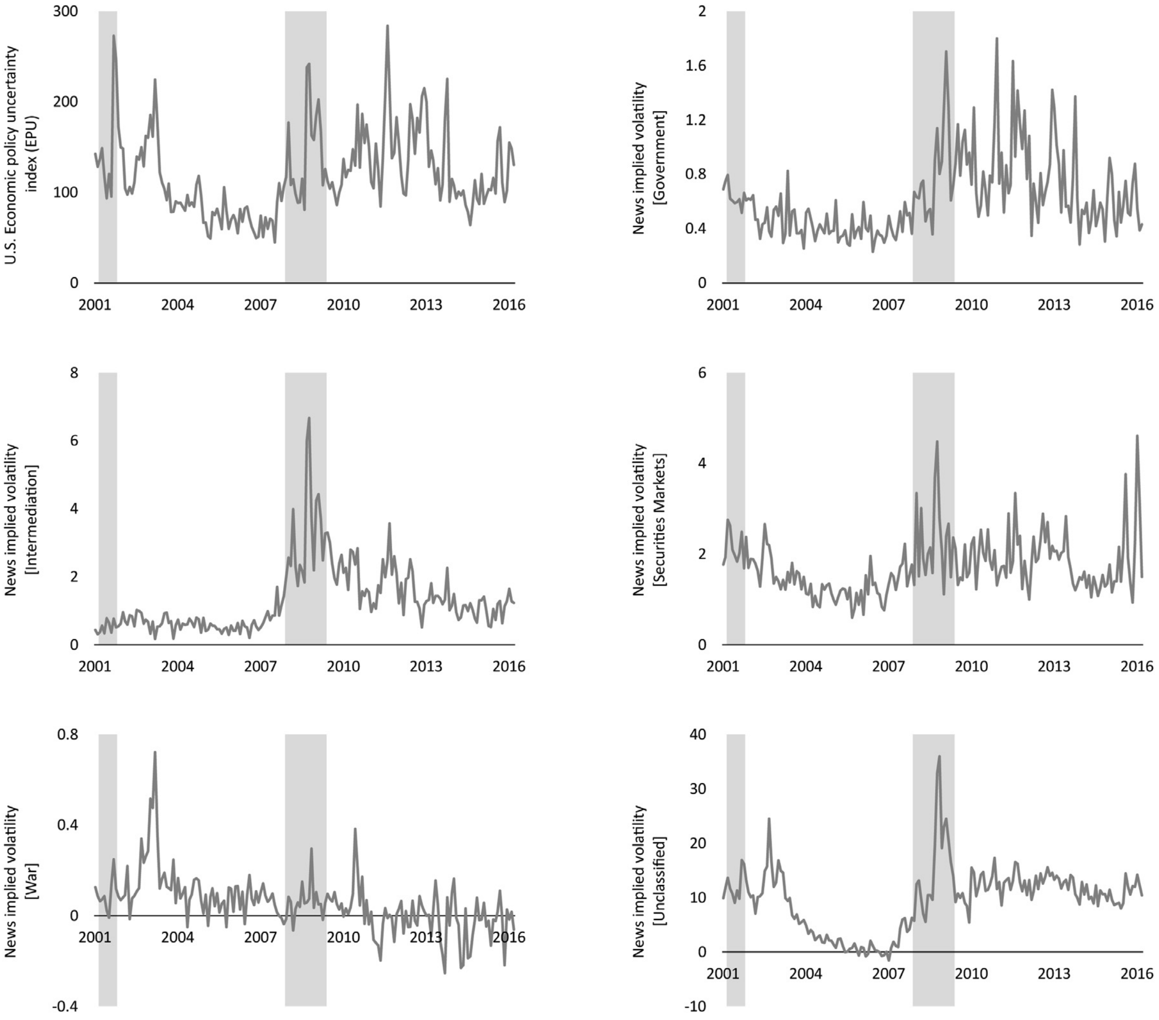

Fig. 2. Economic policy uncertainty (EPU) and categorized news-based implied volatility during the period 2001–2016. Shaded areas represent U.S. recession periods as defined by the National Bureau of Economic Research (NBER).

tion.[9] Moreover, we also yield a better prediction power than when using selected macroeconomic variables instead in model (4). A substantial advantage of text-based news variables over macroeconomic variables is not only due to their superior predictive power, but also that news are available immediately whereas macroeconomic data may be published with a time lag, making news-based variables particularly useful in out-of-time predictions. This reveals the substantial importance of news-based variables for predicting U.S. corporate bond recovery rates.

Five categories of text can be identified as distinct origins of uncertainty which we include in our analysis: government, intermediation, stock markets, war, and unclassified. We observe uncertainty related intermediation having a strongly significant negative impact in models (3) and (5). The most frequent word counts in the intermediation category are the following: "financial", "business", "bank", "credit", and "loan". Intermediation-related news spikes mostly during financial crises and periods of bank failures. Thus, the significantly negative impact on recovery rates observed is in accordance with the intuitive expectation of lower recovery rates during times of financial distress.

The news-related variable of the unclassified category has a significantly negative coefficient in the model specifications (3) and (5). The most frequently occurring words of the unclassified category are "U.S.", "Washington", "gold", "special", and "treasury". The occurrence of the terms "gold" and "treasury" points to 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 unclassified category would explain the significantly negative coefficient of this source of uncertainty.

News-related to the government category is the only news source that has a significantly positive coefficient in our analysis. The most frequently occurring words of this category are "tax", "money", "rates", "government", and "plan". These terms do not necessarily bear a negative connotation. For instance, the prospects of tax cuts or a more expansive fiscal policy might be reflected in news from the government category. So, the expectation of government policies which are perceived as positive 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", and "markets". With uncertainty about financial crises already

[9] In a linear regression analysis not reported here, but whose results are available from the authors, we combine models (2) and (3) of Table 4, controlling for the EPU in the presence of the five news-related measures. In this setting, the EPU becomes insignificant while the variables from government, intermediation and unclassified categories keep their significant relationships with recovery.

Table 4
This table presents the results of the linear regression specifications. The recovery rate of the respective bond is the independent variable. In (1) we use seniority dummies, industry variables, and bond characteristics as independent variables. In (2) we add the economic policy uncertainty (EPU) measure from Baker et al. (2016). In (3), we replace the EPU with news-based measures. In contrast, in (4) we consider the macroeconomic variables selected by stability selection in addition to the seniority dummies, industry variables, and bond characteristics. In (5) we add the combination of news-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)	(4)	(5)
Intercept	44.5654***	47.1636***	38.1582***	46.0907***	33.1257***
	(22.4612)	(18.5472)	(11.8707)	(18.5324)	(8.0126)
EPU		-0.0301*			
		(-1.6564)			
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	0.4179	0.4184	0.4536	0.4462	0.4826
RMSE	26.6211	26.6034	25.7603	25.9202	25.0247
MAE	20.7074	20.6564	19.9213	20.1421	19.2381
AIC	1.96E+04	1.96E+04	1.95E+04	1.95E+04	1.93E+04
BIC	1.97E+04	1.97E+04	1.96E+04	1.96E+04	1.95E+04
Number of bonds	2,079	2,079	2,079	2,079	2,079
Seniority	Yes	Yes	Yes	Yes	Yes
Industry	Yes	Yes	Yes	Yes	Yes
Bond Characteristics	Yes	Yes	Yes	Yes	Yes

reflected through the highly significant intermediation-related uncertainty, we observe the stock market-related news with a negative but insignificant coefficient in models (3) and (5). Further, war-related news has very little variance over our observation period and is not a significant determinant in the linear regression analysis.

Similar to Nazemi & Fabozzi (2018), we select a small number of macroeconomic variables from a large collection of macroeconomic variables, but further compare several different selection techniques for this step as described in Section 5.2. By selecting a small subset of the macroeconomic variables and eliminating the rest of the variables, the outcome model becomes more interpretable compared to using principal component analysis of the macroeconomic variables in Nazemi & Fabozzi (2018). Adding seven selected macroeconomic variables within model (4) presented in Table 4, we achieve a prediction improvement over the basic model specification. Nevertheless, the improvement in prediction accuracy is smaller than achieved by Nazemi et al. (2018). With regards to macroeconomic variables' significance, we observe the following drivers of recovery rates in our analysis. An increase in the inventories-to-sales ratio in the manufacturing industry by 10% coincides with an increase in the recovery rate of 5.0% in model (4). This is contrary to the notion that a lower sales turnover indicates macroeconomic weakness and might lead to a lower recovery rate. Further, in accordance with macroeconomic intuition, the average recovery rate decreases by 1.48% at a time when the number of unemployed civilians rises by 100,000.

Additionally, when inflation of the unit labor cost in the business sector increases by 1%, the average recovery rate decreases by 1.37%. An increase in the 30-year mortgage rate by 1% coincides with a 10% increase of the recovery rate in model (4). Moreover, when the U.S. dollar strengthens by 1% against a trade-weighted basket of foreign currencies, we observe a decrease of 1.2% in the recovery rate.

We provide more evidence on the effects of text-based news variables by examining how they interact with a recession indicator. In Table 5 we present additional regression results including terms for the interaction effects between news-based variables and a recession indicator. Simultaneously, we allow for the respective direct effects of both the recession indicator and the news-based variables. In model (1), we add the recession indicator to a basic specification including seniority dummies, industry variables, and bond characteristics. The recession indicator is a dummy variable indicating recession in the U.S. economy for the periods that are defined as recessions by the National Bureau of Economic Research (NBER). Our dataset covers two recessions according to the NBER definition: March 2001 – November 2001 and December 2007 – June 2009. During these recessions, about 900 of the defaults in our dataset occurred (i.e. 44% of the defaulted corporate bonds). It is worthwhile to mention that while our analysis involving the recession indicator allows us to better understand the economic

Table 5
This table presents the results of the linear regression involving interaction terms between a recession indicator and news-based variables. The recovery rate of the respective bond is the independent variable. In (1) we use seniority dummies, industry variables, and bond characteristics as independent variables, and further add a dummy variable indicating recession in the U.S. economy for periods as defined by the National Bureau of Economic Research (NBER). In (2) we add the news-based variables and account for the interaction between the recession indicator and the news-based variable from the government category. In models (3), (4), (5) and (6) we iterate the interaction terms of the recession indicator and news from intermediation category, securities market category, war category and the unclassfied category, respectively. The interaction terms and their respective singular compontents are highlighted in bold. 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)	(4)	(5)	(6)
Intercept	50.8709***	42.3065***	40.0089***	40.9739***	35.7024***	40.1866***
	(19.7543)	(9.5466)	(9.0664)	(8.0498)	(8.5479)	(9.1934)
Government		**26.7863*****	34.9647***	35.8834***	37.5727***	34.8412***
		(6.3399)	(10.0266)	(10.2786)	(11.0248)	(10.0359)
Intermediation		-8.6486***	**-9.6021*****	-7.3071***	-6.2321***	-7.4705***
		(-5.2494)	(-4.9892)	(-4.4309)	(-3.9725)	(-4.6907)
Securities Market		1.3763	0.6433	**-2.0550**	-0.7461	0.5046
		(0.7789)	(0.3651)	(-1.1095)	(-0.4372)	(0.2902)
War		-11.6967	-12.2428	-13.7096	**-12.9651**	-14.9453
		(-1.2937)	(-1.3512)	(-1.5046)	(-1.3721)	(-1.6430)
Unclassified		-0.2945	-0.4609*	-0.3090	-0.2576	**-0.7210****
		(-1.1815)	(-1.7580)	(-1.2176)	(-0.9500)	(-2.4718)
Recession	-17.0835***	**-26.8014*****	**-19.8319*****	**-25.9045*****	**-13.4593*****	**-23.7072*****
	(-6.3700)	(-6.2730)	(-5.5059)	(-3.4451)	(-4.4292)	(-5.5678)
Government*Recession		**22.2482*****				
		(4.1927)				
Intermediation*Recession			**5.2463*****			
			(2.9068)			
Securities Market*Recession				**6.5848***		
				(1.8361)		
War*Recession					**7.4322**	
					(0.3138)	
Unclassified*Recession						**0.9869*****
						(3.2719)
Adj. R^2	0.4567	0.4921	0.4898	0.4885	0.4877	0.4904
RMSE	25.6678	24.7820	24.8370	24.8677	24.8876	24.8234
MAE	19.7555	18.8180	18.8880	18.9178	18.9596	18.8674
AIC	1.94E+04	1.93E+04	1.93E+04	1.93E+04	1.93E+04	1.93E+04
BIC	1.96E+04	1.95E+04	1.95E+04	1.95E+04	1.95E+04	1.95E+04
Number of bonds	2,079	2,079	2,079	2,079	2,079	2,079
Seniority	Yes	Yes	Yes	Yes	Yes	Yes
Industry	Yes	Yes	Yes	Yes	Yes	Yes
Bond Characteristics	Yes	Yes	Yes	Yes	Yes	Yes

effects of text-based news variables on recovery rates of defaulted U.S. corporate bonds, it cannot be included into out-of-time predictions because it is only fully determined in retrospect to a recession. We observe a significant negative relationship between the recession indicator and recovery. This is in line with the economic intuition that recovery rates are lower during times of distress. We point out that the industry distress dummy variables are still significant and maintain negative coefficients in this specification, hence we conclude that an economy-wide recession operates as a superordinate factor in addition to industry-specific distress factors.

As we have demonstrated the significance of news-based variables in the first part of this section, we now expand our analysis on the effects from text-based news variables on recovery rates of defaulted bonds by iterating interaction terms between text-based news variables and the recession indicator in models (2)–(6) in Table 5. In general, we find that the interaction effects are all positive. Yet, consistent with our previous analysis, news variables and interaction terms related to securities markets and war are insignificant, with the small exception that the interaction between news related to securities markets and recession is significant at the 10% confidence level in model (4). More importantly, we find that the interaction terms in models (2), (3) and (6) are significant with 1% confidence level, i.e. the interactions between recession and those text-based news variables that we found to be significantly related to recovery rates earlier in this section (news from the government, intermediation and unclassified categories). At the same time, the direct effects of news related to government, intermediation and industry distress remain intact in all models. Interestingly, the magnitude of the positive effect from news related to the government in model (2) almost doubles during the recession period as compared to times of economic growth. Apparently, as recovery is expressed by bond prices in our analysis, news related to the government are perceived even more positive by investors during a recession. This is plausible, as substantial actions have been taken by U.S. authorities during the recent global financial crisis in order to mitigate distress, support the economy and provide stability to the financial system. News about these actions will likely lead to increasing investors' confidence.

We also perceive that the positive coefficient of the interaction term of intermediation news and recession (model (3)) has a similar magnitude as the negative coefficient of the direct effect from intermediation news, but is slightly smaller. This shows that during the crisis, intermediation news still negatively affect recovery rates of defaulted U.S. corporate bonds, but to a much smaller extent than during non-recession periods. While the text-based news variable of the unclassified category is significant when including its own interaction term in model (6), it becomes insignificant in most of the other models. Moreover, when comparing the coefficient of the interaction effect of unclassified news with the unclassified news' direct effect, we find that the directional effect of unclassified news on recovery rates turns from negative in times of economic growth to positive during a recession. Although the reasoning behind this is somewhat opaque, it is not implausible given that unclassified news cover the bulk of all news collected, and therefore should include both news that increase concerns but

Table 6
This table presents the results of the linear regression considering only defaulted bonds issued by non-financial firms. The recovery rate of the respective bond is the independent variable. In (1) we use seniority dummies, industry variables, and bond characteristics as independent variables. In (2) we add the economic policy uncertainty (EPU) measure from Baker et al. (2016). In (3), we replace the EPU with news-based measures. 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	37.3171*** (17.8758)	45.3711*** (14.9568)	45.9041*** (13.5024)
EPU		-0.0693*** (-3.6397)	
Government			11.9157*** (3.2147)
Intermediation			-2.4522** (-2.1189)
Securities Markets			-3.6010** (2.0810)
War			-2.3412 (-0.3011)
Unclassified			-0.8354*** (−3.7856)
Adj. R^2	0.3571	0.3649	0.3885
RMSE	25.3207	25.1551	24.6324
MAE	19.3735	19.1727	18.9089
AIC	0.95E+04	0.95E+04	0.95E+04
BIC	0.96E+04	0.96E+04	0.96E+04
Number of bonds	1,020	1,020	1,020
Seniority	Yes	Yes	Yes
Industry	Yes	Yes	Yes
Bond Characteristics	Yes	Yes	Yes

also news that increase the confidence of investors. In summary, the analysis of interaction between news and the recession indicator is an interesting piece of empirical evidence that shows how news can reflect investors' concerns, but are also positively perceived by investors if the news are related to government actions during economic recessions.

We further elaborate whether our observations are rooted in the large fraction of defaulted bonds issued by firms from the financial industry. About 51% of defaulted bonds in our data sample are attributable to the financial industry, hence we remove these bonds and conduct linear regression analysis on the remaining sample, involving 1,020 defaulted bonds that were issued by non-financial firms. In model (1) of Table 6, we consider the basic specification that includes seniority dummies, industry variables, and bond characteristics as independent variables. Model (2) adds the EPU measure and confirms its significant negative relationship to recovery rates that has been previously identified for the whole sample. In model (3), we replace the EPU with text-based news variables and observe significant positive effects of news from the government category, and significant negative effects of news from the intermediation and unclassified categories, which is similar to our findings when involving the whole sample. We find, however, that the coefficients of these variables are substantially smaller as compared to using the whole dataset, indicating that the news variables have a smaller effect on non-financial bonds. As *The Wall Street Journal* is predominantly focused on business and financial news, we are not surprised by this finding. Opposite to involving the whole sample, however, securities markets news become significant, having also a negative effect on recovery rates. In summary, the analysis highlights that government related news keep their unique characteristics, having a significant positive effect on recovery rates of defaulted bonds for both samples including or excluding defaulted bonds issued by financial firms. Although the other text-based news variables consistently have negative coefficients, they appear to be more interchangeable with each other, depending on alternations in the data sample.

Overall, taking into account the significance of three out of the five text-based measures even when controlling for macroeconomic effects points to a time-varying influence of investors' mood on recovery rates. The further increase of prediction performance when combining news-related measures with macroeconomic variables demonstrates that the news-based variables contain additional predictive information compared to macroeconomic variables. This is in line with the finding of Gambetti et al. (2019) when controlling for the business cycle in their analysis of uncertainty measures. Hence, we can conclude that the effect measured by the text-based measures has additional predictive power for recovery rates and is not simply mirroring the already known significance of macroeconomic variables for recovery rate prediction. We further find that news related to the government generally has a unique positive effect on recovery rates of defaulted corporate bonds. The effect magnifies during recession periods, a characteristic that possibly accounts for an increasing confidence among investors, conveyed through the news and reflected in 30-day bond prices. While effects from other news categories tend to be negative, their magnitude decreases during recessions. For unclassified news, the effects even become positive, allowing us to conclude that news does not only have a unidirectional effect on recovery rates of defaulted corporate bonds. We also find that financial news has a smaller effect on recovery rates of defaulted bonds from non-financial issuers.

5.2. Benchmark of variable selection techniques

In this study, we employ machine learning techniques using two different prediction settings. First, we predict out-of-sample. We sort the dataset randomly stratified for the seniority classes. After using 10 folds cross-validation to select the hyperparameters based on the root mean-squared errors (RMSEs) on the training set (70% of the data), we predict out-of-sample on the test set (30% of the data). By following this procedure, we 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 of Bui et al. (2017) by using α=0.5 for a MSE loss when applying their power expectation propagation approach.

In Table 7, we demonstrate that the machine learning techniques outperform traditional statistical techniques during out-of-

Table 7
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 Reg.: 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

sample prediction. Using a random partition of 70% of the dataset as the training set, we 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 for selecting the most important macroeconomic variables. Without regard to the selection technique used for selecting the macroeconomic 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 mean absolute error (MAE). Independent of which selection technique is applied, random forest shows the best predictive out-of-sample performance.

Nazemi & Fabozzi (2018) demonstrate that recovery rate models incorporating LASSO-selected macroeconomic variables outperform those from previous research which include only few macroeconomic variables or principal components. We apply and benchmark three different selection techniques, SparseStep, MC+ and stability selection, for identifying the most relevant macroeconomic variables. For all six prediction techniques, macroeconomic variables selected by SparseStep appear to provide the best predictive accuracy. Thus, finding that selecting the macroeconomic variables with SparseStep instead of LASSO increases predictive accuracy, we yield an improvement on the study from Nazemi & Fabozzi (2018). However, the difference between SparseStep and 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), the predictive accuracy decreases slightly. The regression tree (best RMSE of 22.4956) has the lowest predictive power of the machine learning techniques. Among the traditional approaches, the 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 the linear regression and inverse Gaussian regression techniques. For this reason, we will use SparseStep during out-of-time prediction.

Even though comparability of performance measures across datasets is limited, our results for out-of-sample estimation are on par with the best results in the literature. The lowest RMSE reported by Yao et al. (2015) is 0.2136 for SP LS-SVR during an out-of-sample prediction study. Nazemi & Fabozzi (2018) report the lowest RMSE of 0.1750 for LS-SVR with different intercepts for each seniority class during 10 folds cross-validation. The lowest RMSE during 10 folds cross-validation in the study by Kalotay & 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 (linear regression and inverse Gaussian regression), irrespective of which selection technique is utilized. While this relationship has been documented by Kalotay & Altman (2017); Qi & Zhao (2011); Yao et al. (2015), and Nazemi & Fabozzi (2018), the literature on corporate bonds' recovery rate prediction out-of-time is sparse. In the following, we address this gap in the literature.

5.3. Intertemporal prediction of the recovery rate

Having predicted out-of-sample in the previous section, we now predict out-of-time. We first evaluate the machine learning methods in out-of-time prediction of recovery of portfolios of defaulted bonds in accordance with Kalotay & Altman (2017), and thereafter predict recovery rates on the instrument-level by applying five out-of-time prediction setups inspired by approaches suggested in the asset pricing literature.

5.3.1. Intertemporal prediction of defaulted bond portfolio recovery

As cogently outlined by Kalotay & Altman (2017), prediction out-of-time instead of out-of-sample addresses several issues. Taking into consideration the likelihood of time variation in recovery rates, they point out that reporting forecast performance on a random partition of the dataset is less appropriate. Kalotay & Altman (2017) emphasize the importance of accounting for time variation in recovery rates. In particular, testing performance out-of-time ensures that only sample points observed before the default event were used for training. Further, only investigating performance out-of-time prevents data points from the same issuer and the same exposure to be part of both the training and test set, therefore satisfying the condition that observations in the test set are independent from observations in the training set. We train our models by including data from 2001 until 2011 and use data from the remainder of the sample period (2012 to 2016) as the test set. Following Kalotay & Altman (2017) for ease of comparison, we predict mean recovery rates of portfolios of defaulted bonds. Therefore, 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. We also repeat this analysis moving through time. Starting with the training set from 2002 to 2011, we add another year of data to the training set until we have reached the end of the dataset using training data up to 2014. The bonds from the two years following the training period are used as the 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 8.[10] The bonds from 2001 to 2011 are used as the training set while the bonds from 2012 to 2016 are used as the test set for sampling. Again, machine learning techniques outperform

[10] As we yield the most accurate predictions with SparseStep during out-of-sample prediction, we report only the results 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.

Table 8
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. The best performance measures are highlighted in bold. (IG Reg.: Inverse Gaussian Regression; Lin. Reg.: Linear Regression; Reg. Tree: Regression Tree; SP LS-SVR: Semi-Parametric Least-Squares Support Vector Regression; PEP: Sparse Gaussian Process Approximation with Power Expectation Propagation; RF: Random Forest).

		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 9
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 the test set. So, in the first step we use the bonds from 2001 to 2011 as the 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 the test set. The SparseStep algorithm is used to select the most informative macroeconomic variables. The best performance measures are highlighted in bold. (IG Reg.: Inverse Gaussian Regression; Lin. Reg.: Linear Regression; Reg. Tree: Regression Tree; SP LS-SVR: Semi-Parametric Least-Squares Support Vector Regression; PEP: Sparse Gaussian Process Approximation with Power Expectation Propagation; RF: Random Forest).

		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

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. Interestingly, in the out-of-time prediction setup, the power expectation propagation approach yields the lowest RMSE of 2.6887 while SP LS-SVR (4.2736) and the regression tree (5.1717) exhibit a slightly lower predictive capacity.

In Table 9 we show the out-of-time performance of our models when retraining the models each 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 the test set for sampling. For instance, in the first step we use the bonds from 2001 to 2011 as the training set and sample from the bonds from 2012 and 2013 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 the root mean square error (RMSE) and mean absolute error (MAE), the best performing method in Table 9 is again the power expectation propagation approach with an 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 the lowest percentage deviation among all techniques only for the 1st-percentile, 5th-percentile, and 75th-percentile. In contrast, the 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 the random forest has the lowest percentage deviation for the 90th-percentile (deviating -1.56%).

Table 10
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 bonds from the following two years as the test set. So, in the first step we use the bonds from 2001 to 2011 as training set and the bonds from 2012 and 2013 as the 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. The best performance measures are highlighted in bold. (IG Reg.: Inverse Gaussian Regression; Lin. Reg.: Linear Regression; Reg. Tree: Regression Tree; SP LS-SVR: Semi-Parametric Least-Squares Support Vector Regression; PEP: Sparse Gaussian Process Approximation with Power Expectation Propagation; RF: Random Forest).

			SparseStep					
	# of bonds		IG Reg.	Lin. Reg.	Reg. Tree	SP LS-SVR	PEP	RF
2011	75	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	62	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	91	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	104	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
Mean		RMSE	31.2402	30.7148	13.4058	12.3431	**11.3948**	15.5855
		MAE	29.7744	29.2172	11.0205	10.6562	**9.4109**	13.9478

While Kalotay & Altman (2017) report their lowest RMSE of 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 result is similar for out-of-time prediction when yearly retraining the models. Each of our three best-performing machine learning techniques – the power expectation propagation approach (RMSE = 11.8), the SP LS-SVR (RMSE = 13.2), and the regression tree (RMSE = 13.7) – outperforms their best-performing technique, a mixture model with bagging (RMSE = 15.5). Performing the comparison on a percentile-level, our best techniques outperform the best techniques reported by Kalotay and Altman (2017) for the median and the higher percentiles (75% and 90%) while for the lower percentiles (1%, 5%, 10%, and 25%) the techniques from Kalotay & Altman (2017) are more accurate.

More traditional approaches such as linear regression and inverse Gaussian regression experience significant deterioration during the out-of-time prediction compared with their out-of-sample performance. In contrast, the predictive accuracy of the machine learning techniques such as the newly proposed power expectation propagation approach and SP LS-SVR does not decline when switching from out-of-sample estimation to out-of-time estimation. From these results we conclude that the non-linear relationships between the recovery rate and the explanatory variables are more stable during our observation period than the linear dependencies between the recovery rate and the explanatory variables. In general, we find that the power expectation propagation approach provides the most compelling out-of-time prediction results.

Although we report the average performance measures across all time steps in Table 10, we show 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.

5.3.2. Intertemporal prediction of instrument-level recovery rate

In the following, we focus on intertemporal recovery rate prediction of individual bonds rather than defaulted bond portfolios. We compare the performance of the machine learning methods for individual bonds' recovery rate prediction across five intertemporal prediction setups. We employ intertemporal prediction setups that have been previously used in the asset pricing literature. The standard approach is to divide the data into a subsample for model training, a validation subsample for hyperparameter selection, and a test subsample for prediction performance evaluation. Bali, Goyal, Huang, Jiang, & Wen (2021) use fixed consecutive time windows for training, validation and prediction testing. Bianchi, Bchner, & Tamoni (2020) use an annual rolling prediction window of fixed size that is preceded by a training sample which increases as they move in time, performing cross-validation on the training set for hyperparameter selection. Gu, Kelly, & Xiu (2020) combine both approaches by using a rolling prediction window of fixed size with an increasing training set, however selecting hyperparameters on a fixed-sized validation set which is located between the training and test sets and which they roll forward as they move in time.

First, we train the models on the data 2002 to 2011 while relying on 10-fold cross-validation for tuning hyperparameters, and then predict on the test set 2012 to 2016. This setup is comparable to that used in Table 8 for the portfolio approach. For the subsequent setups, we replace the cross-validation for tuning the hyperparameters with a validation set that is located between the training and the test set. Hence, for the second setup, we split the data into the training set ranging from 2002 to 2010, the year 2011 as the validation set, and the years 2012 to 2016 remain as the test set. Third, we suggest a setup in which we move through time. We start with the years 2002 to 2010 as the training set, 2011 as the validation set and the two subsequent years 2012 and 2013 as the test set. We then move one year in time, increasing the training data by one year but keeping the lengths of the validation and test sets as one year and two years, respectively. The fourth setup is similar to the previous setup except that we also keep the length of the training data fixed and drop older training data, instead of increasing it as we move through time. In the final setup, we apply a daily rolling prediction window that consecutively predicts recovery rates of default-day combinations, i.e. we predict all recovery rates of bonds that defaulted on a given day, and thereafter move on to the next default date in our data on which one or more bonds defaulted. We then retrain the model and predict recovery rates of all bonds which defaulted on that day. For this setup, the validation set considers the 120 most recently defaulted bonds, of which the oldest defaults are added successively to the training set as we move through time, while the last predicted defaulted bonds feed into the validation set.

We denote the training set τ_1, the validation subsample for hyperparameter selection τ_2, and the test set for prediction performance evaluation τ_3. We measure performance on the test sets with RMSE and MAE:

$$\mathrm{RMSE} = \sqrt{\sum\nolimits_{i\in\tau_3} \frac{\left(RR_i - \hat{RR}_i\right)^2}{n}} \tag{1}$$

Table 11
This table shows the performance measures of machine learning methods for five different out-of-time prediction setups. In setup (1), we train the models on the data 2002 to 2011 while relying on 10-fold cross-validation for tuning hyperparameters, and then predict on the test set 2012 to 2016. In (2) we use the year 2011 as the validation set and the years. In (3), we use a rolling prediction window, with 2011 as the validation set and the two subsequent years test set, consecutively moving one year in time, increasing the training data by one year in each step. Setup (4) is similar to setup (3) with the exception of fixed training set length. In setup (5), we apply a daily rolling prediction window that consecutively predicts recovery rates of default-day combinations. For this setup, the validation set considers the 120 most recently defaulted bonds, of which the oldest defaults are added successively to the training set as we move through time. The best performance measures per prediction setup are highlighted in bold. The SparseStep algorithm is used to select the most informative macroeconomic variables. (IG Reg.: Inverse Gaussian Regression; Lin. Reg.: Linear Regression; Reg. Tree: Regression Tree; SP LS-SVR: Semi-Parametric Least-Squares Support Vector Regression; PEP: Sparse Gaussian Process Approximation with Power Expectation Propagation; RF: Random Forest).

Out-of-time prediction setup			IG Reg.	Lin. Reg.	Reg. Tree	SP LS-SVR	PEP	RF
(1)	Fixed window; fixed training length; cross-validation	RMSE	54.3040	56.1158	29.9666	27.6113	**27.2993**	30.2622
		MAE	47.0216	48.2788	23.3972	23.4763	**22.0009**	26.2950
(2)	Fixed window; fixed training length;one year validation	RMSE	55.4472	59.2014	34.3494	**27.9730**	29.0718	31.2301
		MAE	48.7604	52.2000	26.1498	24.4059	**20.7947**	26.8246
(3)	Annual rolling window; increasing training length;one year validation	RMSE	47.5039	47.4405	29.8983	30.7911	**26.2828**	30.7124
		MAE	41.0943	41.2887	22.5964	26.0794	**19.9878**	26.1667
(4)	Annual rolling window; fixed training length;one year validation	RMSE	48.2685	49.6850	32.8556	32.9354	**27.6137**	32.8795
		MAE	41.6908	43.0925	25.7331	28.0902	**22.4910**	28.2036
(5)	Daily rolling window; increasing training length;120 defaults for validation	RMSE	51.1180	52.5641	28.0918	26.7455	**23.7689**	31.0743
		MAE	44.8800	46.6117	21.6381	21.9732	**18.1355**	26.6197

$$\mathrm{MAE} = \sum_{i\in\tau_3} \frac{\left|RR_i - \hat{RR}_i\right|}{n} \tag{2}$$

where $\hat{RR}_i$ is the out-of-time predicted recovery rate and RR_i the actual recovery rate of bond i, and n is the total number of bonds in the test set τ_3.

The results are shown in Table 11. Again, the machine learning methods outperform linear regression and inverse Gaussian regression techniques. Furthermore, consistent with the previous analysis, the power expectation propagation approach performs best and delivers the lowest forecast errors in four of five setups. It yields the best result (RMSE = 23.8) when applied in the setup with a daily rolling prediction window and increasing test set length (Setup (5)). Only in Setup (2), where we apply a fixed prediction window and select hyperparameters during a fixed validation year, SP LS-SVR is performing better than the power expectation propagation approach.

Moreover, we find that for fixed prediction windows, parameter tuning via 10-fold cross-validation across the full historic data (Setup (1)) yields better recovery rate forecasts than using only the last year before the test set as a validation set (Setup (2)). Likewise, the rolling window approach performs better when increasing the training size, considering all historic recovery rate observations (Setup (3)), instead of keeping the length of the training set fixed by adding new training data and dropping old training data while moving through time (Setup (4)). Both of these observations indicate that incorporating the full historic information in calibrating the models is more valuable than calibrating with more recent data. Overall, our analysis demonstrates the benefits of applying the power expectation propagation approach for out-of-time recovery rate prediction.

5.4. Permutation importance of groups of explanatory variables

Here we rank groups of variables to elaborate on the degree of feature importance for recovery rate prediction. We investigate the permutation importance according to Altmann, Toloşi, Sander, & Lengauer (2010) of each group of variables for the performance of the random forest technique in recovery rate prediction. Therefore, we build 11 groups of independent variables as detailed in Internet Appendix A: industry, bond characteristics, seniority, news, and the macroeconomic variables which are separated into 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 examine the importance ranking of groups of variables for the U.S. corporate bonds that defaulted from 2001 to 2016.

Table 12
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

As illustrated in Table 12, bond characteristics are the most important group of variables for recovery rate prediction in our analysis. So, the significance of bond characteristics reported by Jankowitsch et al. (2014) is confirmed by 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, for example, Varma & Cantor (2005) and Jankowitsch et al. (2014). The importance of stock market indicators (ranked third, 14.3448) confirms the significance of the return on the market index reported by Varma & Cantor (2005).

Interestingly, the group of text-based news variables is ranked higher than the widely used industry variables (9.5908 compared with 6.1447), which confirms our findings in Section 5.1 that this group of variables is an important driver of recovery rates. Similarly, the literature has paid little attention to variables indicating international competitiveness which ranked fourth in our analysis with an importance of 13.2206. Having an importance of 2.9321, business cycle variables which include commonly used variables such as GDP growth and the unemployment rate (see, for example, Altman et al., 2005 and Yao et al., 2015) ranked only second to last in our analysis.

Micro-level factors such as the federal funds rate and the term structure reported to be significant by Jankowitsch et al. (2014); Nazemi & Fabozzi (2018) and considered by Qi & Zhao (2011) are ranked seventh with an importance of 4.5070 in our analysis. 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 respectively 2.2154).

The industry of the defaulted bond is reported to be an important determinant of recovery rates by Altman & Kishore (1996). Further, Acharya et al. (2007) introduce two industry distress dummy variables indicating a 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. The ranking groups of variables provides insights into which groups of covariates have more information for predicting recovery rates for corporate bonds. Interestingly, groups of variables involving text-based news or international competitiveness, which have been neglected by previous research, have higher importance ranks than industry-factors or macroeconomic variables that have previously been extensively studied by researchers. The finding suggests that these unexplored higher ranking groups of variables provide potentially promising fields of research on the economic mechanisms in recovery rate determination.

6. Conclusions

The recovery rate is a key risk parameter in credit risk. Though there is substantial literature on out-of-sample recovery rate estimation for corporate bonds, most approaches employed suffer from two main shortcomings. The assumption of a time-invariant recovery rate distribution is unrealistic. Moreover, assuming the independence of samples when multiple defaulted bonds from the same issuer are part of both training and test set results in 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 prediction of recovery rates for defaulted U.S. corporate bonds over the period 2001–2016 in several intertemporal setups 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 prediction. Employing semiparametric least-squares support vector regression, a power expectation propagation approach, regression tree, and random forest yields significantly higher predictive out-of-time accuracy than the traditional statistical techniques. In particular, the newly proposed power expectation propagation approach achieves the most compelling prediction results under several different out-of-time prediction setups. Interestingly, we also find that out-of-time prediction accuracy benefits from considering a longer history of data for model generation, rather than merely using more recent data and dropping older data points. We test whether news-implied measures and its five components can predict the recovery rates of corporate bonds. These measures relied on machine learning techniques to uncover information from the front-page coverage of *The Wall Street Journal*. Interestingly, we find that investors' uncertainty about the government, intermediation, and the economy are significant drivers of recovery rates. Government-related news are associated with higher recovery rates, especially during recessions. News that are generally negatively associated with recovery rates tend to be less harmful or even turn out to be supportive for recovery rates in times of economic downturns. We further discover that recoveries of bonds issued by non-financial firms are less impacted by financial news.

We benchmark three techniques for selecting the most informative macroeconomic factors from a broad range of macroeconomic variables. Among the selection techniques examined, the SparseStep algorithm selects those macroeconomic variables which contribute the most to recovery rate prediction accuracy. Lastly, studying the permutation importance of the groups of 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. However, groups of variables involving text-based news or international competitiveness, which have drawn little or no attention in previous research, appear to be more important in explaining recovery rates than previous studies would suggest.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ejor.2021.06.047

References

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

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

Altman, E. I., & 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 (Oxford, England), 26*(10), 1340–1347.

Baker, S. R., Bloom, N., & Davis, S. J. (2016). Measuring economic policy uncertainty. *Quarterly Journal of Economics, 131*(4), 1593–1636.

Bali, T. G., Goyal, A., Huang, D., Jiang, F., & Wen, Q. (2021). Different strokes: Return predictability across stocks and bonds with machine learning and big data. *Swiss Finance Institute, Research Paper Series*, 20–110.

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

Betz, J., Kellner, R., & Rsch, D. (2021). Time matters: How default resolution times impact final loss rates. *Journal of the Royal Statistical Society, Series C (Applied Statistics), forthcoming*.

Bianchi, D., Bchner, M., & Tamoni, A. (2020). Bond risk premiums with machine learning. *Review of Financial Studies, 34*(2), 1046–1089.

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.

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

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.

Cortes, G. S., & Weidenmier, M. D. (2019). Stock volatility and the great depression. *Review of Financial Studies, 32*(9), 3544–3570.

Donovan, J., Frankel, R. M., & Martin, X. (2015). Accounting conservatism and creditor recovery rate. *Accounting Review, 90*(6), 2267–2303.

Doshi, H., Elkamhi, R., & Ornthanalai, C. (2018). The term structure of expected recovery rates. *Journal of Financial and Quantitative Analysis, 53*(6), 2619–2661.

Frye, J. (2000). Depressing recoveries. *Risk (Concord, NH)*, 106–111.

Gambetti, P., Gauthier, G., & Vrins, F. (2019). Recovery rates: Uncertainty certainly matters. *Journal of Banking & Finance, 106*, 371–383.

Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *Review of Financial Studies, 33*(5), 2223–2273.

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

Hurlin, C., Leymarie, J., & Patin, A. (2018). Loss functions for loss given default model comparison. *European Journal of Operational Research, 268*(1), 348–360.

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., & Subrahmanyam, M. G. (2014). The determinants of recovery rates in the US corporate bond market. *Journal of Financial Economics, 114*(1), 155–177.

Jansen, J., Das, S. R., & Fabozzi, F. J. (2018). Local volatility and the recovery rate of credit default swaps. *Journal of Economic Dynamics and Control, 92*, 1–29.

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

Khieu, H. D., Mullineaux, D. J., & Yi, H. C. (2012). The determinants of bank loan recovery rates. *Journal of Banking & Finance, 36*, 923–933.

Kim, H., & Kung, H. (2016). The asset redeployability channel: How uncertainty affects corporate investment. *Review of Financial Studies, 30*(1), 245–280.

Krüger, S., & Rösch, D. (2017). Downturn LGD modeling using quantile regression. *Journal of Banking & Finance, 79*, 42–56.

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

Metz, A., Sorensen, S., Keisman, D., & Chiu, D. (2012). Trading prices as predictors of ultimate corporate recovery rates. *Moodys Investor Service, Special Comment*.

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

Nazemi, A., & Baumann, F. (2021). Corporate bond recovery rate and financial markets. *Working paper*.

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

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

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

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

Schläfer, T., & Uhrig-Homburg, M. (2014). Is recovery risk priced? *Journal of Banking & Finance, 40*, 257–270.

Varma, P., & Cantor, R. (2005). Determinants of recovery rates on defaulted bonds and loans for north american corporate issuers:1983-2003. *Journal of Fixed Income, 14*(4), 29–44.

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 support vector machines. *European Journal of Operational Research, 263*(2), 679–689.