Interpretable Machine Learning for Creditor Recovery Rates

Abdolreza Nazemi

Karlsruhe Institute of Technology, Karlsruhe, Germany, email: abdolreza.nazemi@kit.edu

Jonas Rauch

Karlsruhe Institute of Technology, Karlsruhe, Germany, email: jonas.rauch@student.kit.edu

Frank J. Fabozzi

EDHEC Business School, Nice, France, email: frank.fabozzi@edhec.edu

Abstract

Machine learning methods have achieved great success in modeling complex patterns in finance such as asset pricing and credit risk that enable them to outperform statistical models. In addition to the predictive accuracy of machine learning methods, the ability to interpret what a model has learned is crucial in the finance industry. We address this big challenge by adapting interpretable machine learning to the context of corporate bond recovery rates modeling. In addition to the best performance, we show the value of interpretable machine learning by finding drivers of recovery rate and their relationship that cannot be discovered by the use of traditional machine learning methods. Our findings are financially meaningful and also consistent with the findings in the existing credit risk literature.

Keywords: Interpretable machine learning, risk management, recovery rate, corporate bonds

1 Introduction

In recent years, financial economists have been attracted to employ machine learning models for the canonical problems in finance. Many financial decision-makers require models that can find which variables could potentially be relevant and the exact relationship between predictors and the target variable. Most of the machine learning models are black boxes that cannot interpret their results in a way that economists, managers, and investors can understand. In this paper, we implement interpretable machine learning models for recovery rate determinants of defaulted corporate bonds. This paper shows that it is not necessarily true the accuracy–interpretability trade-off in machine learning models.

The determination of creditors' recovery rates in the event of default according to Basel II and III as a key risk parameter is important. Pillar I of the Basel Accord II and III allows financial institutions to use their own credit risk estimates. Consequently, accurate and reliable approaches to estimating recovery rates are needed. Many recent studies have proposed and tested machine learning models to predict recovery rates, these studies show that machine learning methods outperform traditional statistical methods for predicting recovery rates of corporate bonds, personal loans, and leasing (see, for example, Qi & Zhao (2011), Hartmann-Wendels et al. (2014), Nazemi & Fabozzi (2018), Hurlin et al. (2018), Nazemi et al. (2022)). Under many regulations, Banks should be able to interpret their risk models. Accordingly, it is equally important to be able to explain the results of the machine learning models used.

In this regard, a common belief is that models face a trade-off between interpretability and accuracy. Therefore, some works propose to explain complex but accurate black-box models. In this case, however, one cannot look inside the particular model and explain it directly. Only with explainable artificial intelligence methods is it possible to estimate how a model generates its results, which features are important or unimportant, and what is the influence of certain feature values on the result. Rudin (2019) argues that an explanation of highly complex black-box models is not necessary, but that there are also intrinsically interpretable models with almost as good accuracy. These can be interpreted directly and thus provide added value. In this paper, we examine such models by applying them to the prediction of creditor recovery rates and interpreting the models as well as their the results. In addition, the models are compared with standard black-box models in terms of performance and interpretability.

Our primary contribution to the recovery rate literature is to address the above important shortcoming by developing direct interpretable machine learning models for adding interpretability to risk models. To the best of our knowledge, our paper is the first to apply direct interpretable machine learning models for

recovery rate prediction even in empirical finance. We compare the intertemporal prediction performance of interpretable machine learning models with parametric and non-parametric techniques across out-of-time prediction setup. Our findings show that interpretable machine learning models significantly outperform parametric and traditional machine learning methods for predicting recovery rates of corporate bonds. The best out-of-time prediction accuracy is achieved using an interpretable machine learning model, outperforming traditional machine learning models by Qi & Zhao (2011), Hartmann-Wendels et al. (2014), and Nazemi & Fabozzi (2018) as well as machine learning models by other studies such as Yao et al. (2015), Nazemi et al. (2018), and Nazemi et al. (2022). The relations between the recovery rates and variables that we found are consistent with the recovery rate literature and are economically meaningful. More importantly, our model is globally interpretable, the relationships between each explanatory variable and the recovery rate are interpretable, and can be easily explained to managers and investors. Our model can find the most important risk divers and case-based explanations.

We divided our work as follows: In Section 2, approaches for creditor recovery rate estimation are reviewed and in Section 3, we present existing interpretable machine learning methods are and explain their functionality. Section 4 addresses with the empirical analysis of the dataset used and Section 5 presents the results and presents prediction accuracy and model interpretations. 6 provides our conclusions.

2 Literature Review

Altman & Kishore (1996) show that recovery rates depend strongly on the seniority and industry sector of the issuer. They show a higher average recovery rate with higher seniority and higher collateral security. Utilities and chemical companies have significantly higher recovery rates, and this is independent of seniority. Schuermann (2004) confirms this for the utility sector and attributes a higher recovery rate to technology and telecommunications companies. Varma & Cantor (2005) state that macroeconomic variables are important in estimating recovery rates. Altman et al. (2005) argues that default rates, seniority, and collateral levels are important determinants of corporate bond recovery rates. Acharya et al. (2007) examine industry effects and conclude that general industries that are in trouble have lower recovery rates. Chava et al. (2011), Jankowitsch et al. (2014), Mora (2015), and Nazemi & Fabozzi (2018) use a range of macroeconomic variables for the determinants of recovery rates in the US corporate bond market. Mora (2015) states that macroeconomic effects do not have the same impact in every industry. Nazemi & Fabozzi (2018) report that the recovery prediction with LASSO-selected macroeconomic variables from a large number of

macroeconomic variables outperform suggested models in the literature with a few macroeconomic variables. However, Jankowitsch et al. (2014) also introduce liquidity measures in addition to a comprehensive set of bond characteristics, firm fundamentals, and macroeconomic variables as a new driver of recovery rates. Gambetti et al. (2019) note the importance of economic uncertainty to explain a large fraction of the systematic variation of recovery rates. Nazemi et al. (2022) use news-based variables taken from front-page articles in The Wall Street Journal, they show the impact of the news-implied measures on recovery rate.

Machine learning methods are increasingly used for predicting recovery rates usually outperforming traditional methods. Qi & Zhao (2011) use various parametric and nonparametric models to estimate recovery rates. The nonparametric methods such as regression trees and neural networks performed better than the parametric methods. Altman & Kalotay (2014) show that mixtures of Gaussian distributions based on the characteristics of the debtor, the instrument, and the credit market conditions perform better than parametric regression models in out-of-time prediction. Bastos (2014) uses ensembles of regression models that perform better than a single model. Yao et al. (2015) use Support Vector Regression (SVR), among others, to predict recovery rates of U.S. corporate bonds and show that they perform better than conventional methods. Nazemi et al. (2017) add principal components derived from 104 macroeconomic variables to their regression. Fuzzy decision fusion techniques improved predictive accuracy. The implementation of principal components from over 100 economic variables was combined with a multi-factorial framework using utilization instruments and industry variables by Nazemi et al. (2018). This resulted in better accuracy of the Least-Squares SVR methods and linear regression used.

Nazemi & Fabozzi (2018) compare linear regression, SVRs, tree approaches, least absolute shrinkage and selection operator (LASSO), and ridge regression. The models that include the macroeconomic variables selected by LASSO outperform the models that include few macroeconomic variables. Most studies applied out-of-sample or in-sample settings to predict and evaluate the recovery rates of corporate bonds. Kalotay & Altman (2017) highlight the serious shortcomings of the k-fold cross-validation method (out-of-sample setting) for evaluating the predictive accuracy for the recovery rate of corporate bonds. Nazemi et al. (2022) compare several machine learning models such as inverse Gaussian regression, linear regression, regression trees, random forest, semi-parametric least-squares SVR, and introduce sparse Gaussian process approximation with power expectation propagation for out-of-sample and out-of-time predictions. They find that their power expectation propagation approach, as a machine learning approach, provides the most compelling prediction results. Nevertheless, characteristics of bonds, seniority, or stock market indicators are ranked as more important than news-based indicators. Table 1 summarizes studies on recovery rate

modeling for U.S. corporate debt instruments and the performance of machine learning models.

Besides these studies the recovery rates of corporate bonds, a good number of studies[1] apply machine learning models for recovery rates for loans and personal credits. All of these prior studies of recovery rates report machine learning models' performance advantage over parametric models. Machine learning methods have outperformed statistical models in different areas of finance.

In our study, we compare the out-of-time prediction power of our suggested interpretable machine learning models with the power expectation propagation, linear regression, SVR, tree approaches as a benchmark. Many of the models and approaches mentioned here provide very good results in recovery rate prediction. However, the disadvantage of most of the accurate methods mentioned here is that they are difficult to interpret. Especially with more complex models this becomes clear. As explainability of the models becomes more prominent, Bastos & Matos (2022) compare two explainable glass-box models (Fractional Regression and Decision Trees) with a black-box model (gradient boosting). They perform out-of-sample and out-of-time recovery rate prediction for the Moody's Ultimate Recovery Database. With the help of Shapley Additive Explanations (SHAP), feature importance was determined while feature influences were explained using Acculturated Local Effects (ALE) plots. They found that the main determinants of recovery rates under the black-box model differ from those under the parametric glass-box model. They claim that a black-box model requires more effort to explain, but that it should be preferred to glass-box models because they are more accurate. Therefore, Bastos & Matos (2022) suggest using black-box models and explaining them because of their better prediction performance. Kellner et al. (2022) propose a novel feature importance measure to explain neural networks for default prediction. All these approaches have in common that they refer to black-box models or explaining well performing black-box models very recently, but not to finding interpretable machine learning models that perform as well or better as suggested by Rudin (2019). The direct interpretation of machine learning models has crucial advantages over explaining them. Rudin (2019) provides several reasons why interpretable models should be preferred over explaining black-box models. In contrast to Kellner et al. (2022), and Bastos & Matos (2022), we compare the out-of-time prediction power of our suggested direct interpretable machine learning models with the power expectation propagation, linear regression, SVR, and tree approaches as a benchmark in our study.

Interpretability is the extent to which a person can understand or consistently predict the reason for a particular decision made by a machine learning model. A distinction is made between intrinsic and post-

[1] Hartmann-Wendels et al. (2014), Yao et al. (2017), Hurlin et al. (2018), Kaposty et al. (2020), Bellotti et al. (2021).

hoc interpretability or explainable artificial intelligence (XAI). Intrinsic interpretability is associated with a reduction in the complexity of the model, which leads to higher interpretability and is thus a model property. Interpretations can be categorized by their scope, which refers to the part of the prediction process they are intended to explain. There are several possibilities when interpreting machine learning models. Murdoch et al. (2019) gives a good overview of these. Global interpretability means how the model makes decisions. This can be done by better understanding the data features and the individual components learned. The importance of certain features can be examined (i.e., how relevant are the features to the model's decision). Similarly, by looking at individual features, their influence on the outcome can be investigated. In addition, models can also be explained locally. In this case, an individual data point or a group of data points are evaluated to explain the prediction results of the model. This can be useful, for example, when investigating a misprediction. Post-hoc interpretability or XAI involves analyzing black-box models after training and attempting to explain how these models derive their results, as stated by Murdoch et al. (2019).

Direct interpretation of machine learning models has key advantages over their explanation. Rudin (2019) cites several reasons why interpretable models should be preferred over explanatory black-box models. In particular, XAI methods provide explanations that do not adhere to the computations of the original model. Therefore, the explanations must always be wrong because they do not fully match the original model. If the explanations were completely correct, they would be faithful to the original model and the model would no longer be necessary. This then corresponds exactly to the case of an interpretable model. Thus, it can be assumed that any explanation method for a black-box model can be an inaccurate representation of the original model in parts of the feature space. Similarly, according to Rudin (2019), explainable methods do not make sense or do not provide enough detail to understand what the black-box is doing. For this reason, we investigate intrinsically interpretable machine learning models for regression tasks to determine creditor recovery rates. These methods can be understood by users without further explanation methods. At the same time, this work is more concerned with the qualitative impact of features on the overall outcome, which has received less attention in the work presented.

Table 1: Overview of recovery rate models for U.S. corporate debt instruments in literature

Author(s)	Data	Period	Method(s)	Machine learning	ML outperform
Frye et al. (2000)	Corporate bonds	1983-1997	Conditional model	No	-
Altman et al. (2005)	Corporate bonds	1982-2002	Multivariate, logistic, logarithmic and linear regression	No	-
Varma & Cantor (2005)	Bonds and loans	1983-2003	Multivariate regression	No	-
Acharya et al. (2007)	Bonds and loans	1982-1999	Multivariate regression	No	-
Bruche & González-Aguado (2010)	Bonds	1974-2005	Markov chain	No	-
Chava et al. (2011)	Moody's Ultimate Recovery	1980-2008	Linear, logit and probit	No	-
Jacobs & Karagozoglu (2011)	Moody's Ultimate Recovery	1985-2008	Beta-link generalized linear model	No	-
Qi & Zhao (2011)	Moody's Ultimate Recovery	1985-2008	Regression, fractional response, regression, transformation regression tree, neural network	Yes	Yes
Altman & Kalotay (2014)	Moody's Ultimate Recovery	1987-2011	Parametric regressions, regression trees, Bayesian conditional mixture	No	-
Jankowitsch et al. (2014)	Corporate bonds	2002-2010	Multivariate regression	No	-
Donovan et al. (2015)	Moody's Ultimate Recovery	1994-2011	Univariate and multivariate regression	No	-
Mora (2015)	Moody's Default Risk	1970-2008	Univariate and multivariate regression	No	-
Yao et al. (2015)	Moody's Ultimate Recovery	1985-2012	Linear regression, fractional response regression, SVRs, two-stage model	Yes	Yes
Kalotay & Altman (2017)	Moody's Ultimate Recovery	1987-2011	Inverse Gaussian regressions, mixture models, regression trees	Yes	No
Nazemi & Fabozzi (2018)	Corporate bonds	2002-2012	Linear regression, SVRs, bagging, boosting, LASSO, ridge	Yes	Yes
Gambetti et al. (2019)	Moody's Recovery	1990-2013	Beta regression	No	-
Wang et al. (2020)	Moody's Ultimate Recovery	1987-2015	LASSO, Bayesian estimation	No	-
Sopitpongstorn et al. (2021)	Moody's Ultimate Recovery	1994-2012	Local logit regression, Bounded parametric regression	No	-
Nazemi et al. (2022)	Corporate bonds	2001-2016	Regression, SVRs, Inverse Gaussian regression, random forest, power expectation propagation	Yes	Yes
Bastos & Matos (2022)	Moody's Ultimate Recovery	1987 - 2010	Fractional regression, decision trees, gradient boosting	Yes	Yes

3 Interpretable Machine Learning Methods

In this section, we describe the collection of interpretable machine learning methods that we use in our work. We present two approaches and explain them briefly in mathematical background, their functionality and in particular how these models can be interpreted.

3.1 Explainable Boosting Machine

The Explainable Boosting Machine (EBM) proposed by Nori et al. (2019) is a tree-based gradient boosting model built on a Generalized Additive Model (GAM) that was first introduced by Hastie & Tibshirani (1986). A GAM is a linear model that is also suitable for learning non-linear features. The relationships between arguments and target values are described by learned functions. By adding up all the individual relationships, it is possible to make the final prediction. This can be determined as follows:

$$g(E(y)) = \beta_0 + \sum_{i=1} f_i(x_i) = \beta_0 + f_1(x_1) + f_2(x_2) + \ldots + f_m(x_m) \tag{1}$$

where y is the predicted recovery rate for corporate bonds, $E(y)$ is the expected value, g is the linking function linking the expected value to the explanatory variables $x_1, \ldots, x_m$ and f_i are nonlinear smooth functions learned by the model for each single variable. The advantage of using GAMs is that the estimates are easy to interpret and the performance is better than standard linear regression because it handles nonlinear relationships. An EBM is based on GAMs but has some improvements. It mitigates the effects of co-linearity and learns the best feature function for each feature using tree-based methods. Moreover, pairwise interactions can also be detected, which increases performance but maintains interpretability, as shown by Lou et al. (2013). Therefore, the EBM has the form of:

$$g(E(y)) = \beta_0 + \sum_{i} f_i(x_i) + \sum_{i,j} f_{i,j}(x_i, x_j). \tag{2}$$

Training EBMs is a combination of bagging and boosting. A small tree is trained for each feature, while the trees grow by traversing the features multiple times.

The interpretability of this model can be described as follows. By examining f_i or $f_{i,j}$, one can understand the contribution of each variable or pair of variables to the recovery rate. We can either examine the meaning of the various input variables, but equally, by looking at the concrete shape of the function, we can observe how the output of each summand is derived from its inputs. Thus, the EBM is very intelligible. An

implementation of the EBM can be found in the InterpretML framework by Nori et al. (2019), that we also used for our work.

3.2 Neural Additive Models

The Neural Additive Models (NAMs) introduced by Agarwal et al. (2021) combine the strengths of neural networks with the inherent understandability of GAMs. NAMs learn a linear combination of neural networks, each trained on one variable. The networks are trained using backpropagation and can learn any relationship between their input feature and their output. The outputs of all networks are used as inputs to a GAM. Therefore, the output can be denoted as:

$$g(E(y)) = \beta_0 + \sum_{i=1}^{m} h_i(x_i) = \beta_0 + h_1(x_1) + h_2(x_2) + \ldots + h_m(x_m). \tag{3}$$

The equation has the same form as a normal GAM while h_i for $i = 1, 2, \ldots, m$ are not functions of a pre-defined form but rather the neural networks' input-output mapping functions. A sample structure of a NAM is shown in Figure 1. The interpretability of the NAM is similar to the EBM in that by looking

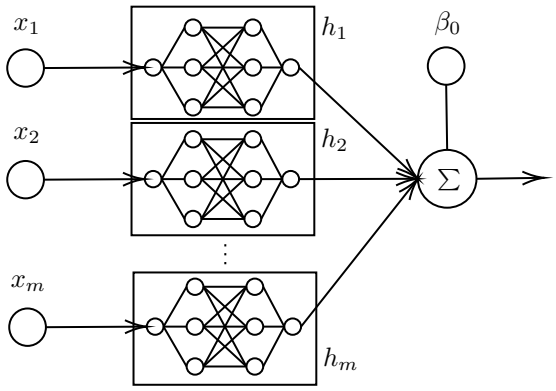

Figure 1: General structure of a NAM. Each feature is input in a neural network that learns a function $h_i(x_i), i \in \{1, \ldots, m\}$. A biased linear combination of the networks' outputs results in the output of the NAM.

at the parameterization functions h_i with $i = 1, \ldots, m$, the influence of each feature on the output can be determined. It is irrelevant how the neural networks obtain these functions, since for each network only one feature is used as input and thus depends only on this feature. Agarwal et al. (2021) show that a NAM for the FICO (2018) (Fair Isaac and Company) dataset significantly outperforms linear regression, i.e., has a significantly lower mean square error for the regression task. It gives the same performance as EBM, while XGBoost and using a neural network give better results, however the latter two methods are less

interpretable. Similar approaches to the NAM are also proposed by Yang et al. (2021) and Vaughan et al. (2018).

4 Data Description

The dataset is drawn from a variety of sources including S&P Capital IQ, Bloomberg, and the Federal Reserve Bank of St. Louis. Our sample consists of 2,080 U.S. corporate bonds that defaulted between 2001 and 2016, drawn from the S&P Capital IQ (Capital IQ) database. Bond issuers are assigned to various industries, such as industrials, consumer cyclical, consumer staples, telecommunications, commodities, utilities, energy, financial services, and information technology. We also consider whether an industry is in distress and incorporate various seniority aspects of the bonds. In addition, the Federal Reserve Bank of St. Louis database (FRED, Federal Reserve Economic Data) is used along with aggregate default data from Fitch Ratings to retrieve macroeconomic variables.[2] We categorized explanatory variables into industry indicators, seniorities, company variables, bond characteristics, and macroeconomic variables.

The density and frequency of recovery rates in our dataset are shown in Figure 2. Most recovery rates have a value of 0% to 10%. There is also another peak at 60% to 70% and a few values above 100%. No typical distribution such as normal, uniform, or bimodal distribution can be observed.

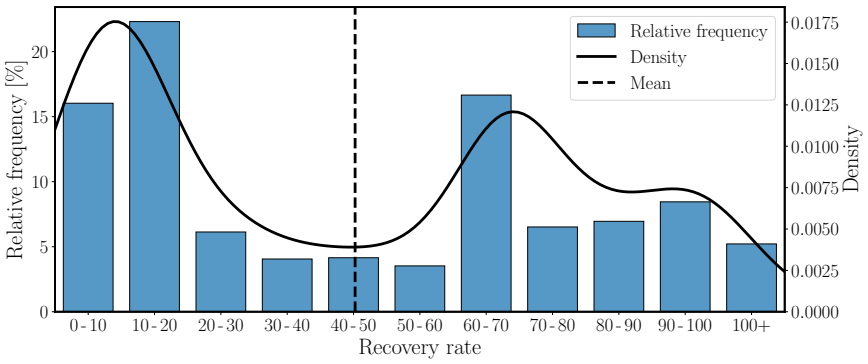

Figure 2: Relative distribution and density of recovery rates in dataset.

Figure 3 shows the recovery rates of U.S. corporate bond over time. The early 2000s recession as a result of the dot-com bubble and the consequences of the 2007-2008 financial crisis are clearly visible. Other than that, the recovery rate values are relatively evenly distributed. To compare and evaluate the output of the interpretable models, the variables' correlation in the dataset with the recovery rate should be

[2] In this study, we use the same dataset as Nazemi et al. (2022). A full description can be found in their paper.

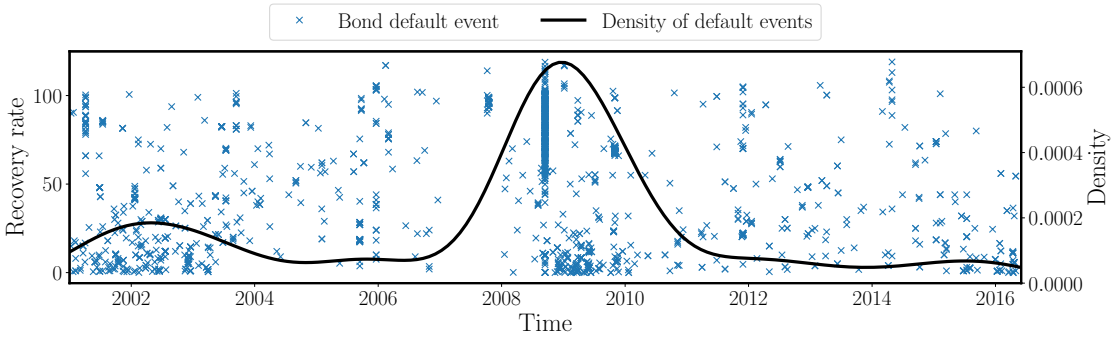

Figure 3: Relative distribution of bond defaults in the dataset. Each cross represents one bond default event.

investigated. Figure 4 shows this relationship for the 17 variables most strongly correlated with the recovery rate. By looking at the correlations between the variables and the recovery rate, it is possible to draw some conclusions about which variables might be important for predicting the recovery rate and how the prediction is influenced by those variables. A machine learning model could also learn these correlations and use them to perform predictions. The financial industry dummy variable shows a strong positive correlation with the recovery rate, and zero-coupon bonds are also correlated with high recovery rates. Secured bonds correlate positively, while subordinated seniority shows a negative correlation. Convertible bonds, bonds from the communications sector, and consumer goods sector may tend to have a lower recovery rate, as they are also negatively correlated. The same applies to industries that are under distress. A high manufacturers inventories to sales ratio also correlates negatively with the recovery rate. A complete correlation matrix of the selected variables can be found in Appendix B. We will compare the results of the correlation analysis with the results of the interpretation of the models to validate our interpretations.

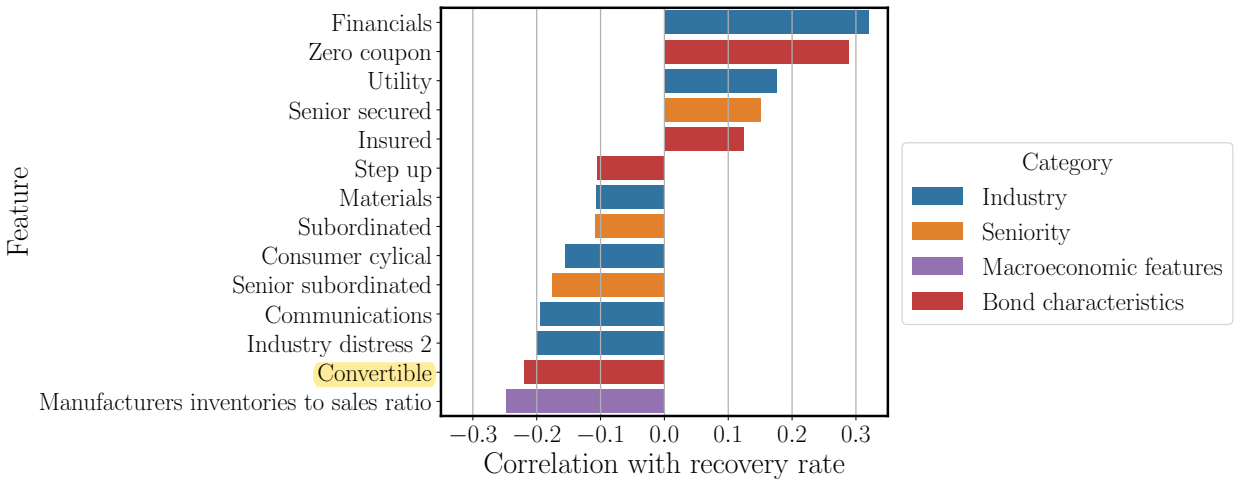

Figure 4: Correlation of variables with recovery rate. Categories of each variable is represented with its bar color. Correlations over ±0.1 are represented to indicate which variables are highly positive or highly negative correlated with the recovery rate.

5 Empirical Results

The models reviewed in Section 3 were trained on the dataset described in Section 4. In the following, we will discuss the individual interpretable machine learning models' accuracy and then present the model interpretation results by measuring variable importance and influences. We will express the advantage of the models used compared to conventional methods for bond recovery prediction.

5.1 Out-of-time Prediction Performance

Kalotay & Altman (2017), and Nazemi et al. (2022) state that only out-of-time prediction is feasible in real-world applications of corporate bond recovery rate prediction. They argue that out-of-sample prediction assumes that there is time invariance in the data-generating process, which leads to a look-ahead bias in the predictions. Further, out-of-sample prediction is implicitly based on the equivocal hypothesis that there is an independency of recovery rates of two defaulted bonds issued by the same company. However, only out-of-time estimation can guarantee that multiple bonds of the same company that defaulted at the same time do not appear in both training and test datasets. To compare the results of the direct interpretable machine learning models with the results of the traditional machine learning and statistical models for predicting the recovery rates of U.S. corporate bonds, we conducted two out-of-time experiments with the training, validation, and testing splits done in Nazemi et al. (2022).

In out-of-time prediction, in contrast to out-of-sample prediction, the training test split is realized via the bond's temporal components. That is, bonds that have failed up to a certain point in time in the past are used as the training dataset. Defaulted bonds of the recent past after this point in time are used as the test dataset. With out-of-sample prediction, bonds with the same default dates can be in both the training and test datasets, since the split is made randomly. For our first experiment, we use bonds from 2001 to 2011 for training and validation, and test the models with bonds from 2012 to 2016. The hyperparameters of NAM and EBM were optimized by 10-fold cross-validation. In the following, we compare the accuracy of the models with Nazemi et al. (2022) results. Therefore, we use the metrics root mean square error (RMSE):

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} e_i^2} \tag{4}$$

and mean absolute error (MAE):

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |e_i| \tag{5}$$

Table 2: Results of out-of-time prediction. 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 for validation purposes. (G. boost.: Gradient boosting, EBM: Explainable Boosting Machine, Lin. reg.: Linear regression; Reg. tree: Regression tree; PEP: Sparse Gaussian process approximation with Power Expectation Propagation; SP LS-SVR: Semi-Parametric Least-Squares Support Vector Regression)

Out-of-time prediction setup			G. boost.	Lin. reg.†	Reg. tree†	PEP†	SP LS-SVR †	**EBM**	**NAM**
(1)	Fixed window; fixed training length; cross-validation	RMSE	27.941	56.116	29.967	27.299	27.611	**26.790**	**23.824**
		MAE	23.752	48.279	23.397	22.001	23.476	**22.343**	**19.737**
(2)	Fixed window; fixed training length; one year validation	RMSE	32.659	59.201	34.349	29.072	27.973	**29.412**	**25.826**
		MAE	25.924	52.200	26.150	20.795	24.406	**23.365**	**20.345**

† Values obtained from Nazemi et al. (2022)

The deviation between the actual and predicted value of the out-of-time bond recovery regression sample is calculated by $e_i = RR_{i,actual} - RR_{i,forecast}$ and n denotes the number of samples. The results are shown in Table 2. Nazemi et al. (2022) found that the sparse Gaussian process approximation with Power Expectation Propagation (PEP) yielded the lowest errors in this out-of-time prediction scenario, outperforming standard methods such as linear regression or regression tree. While EBM was able to achieve similar results for MAE and RMSE, NAM outperforms all other methods.

5.2 Variable Importance

As shown, the models used outperform suggested models in the literature on corporate bonds' recovery rate out-of-time prediction. Another advantage of these methods besides a very good recovery rate prediction is their interpretability, which is missing in most literature dealing with recovery prediction, as Hartmann-Wendels et al. (2014), Kalotay & Altman (2017), Yao et al. (2017), Nazemi & Fabozzi (2018), Hurlin et al. (2018), Kaposty et al. (2020), and Nazemi et al. (2022). Therefore we will present the results of the model's interpretation in the following.

First, we will look at global variable importance. Dataset-level variable importance values attempt to capture how strongly individual variables in a dataset contribute to a prediction. These values can provide insight into which variables the model has identified as important for particular outcomes. This is similar to correlation analysis, but refers to relationships learned by the model that go beyond simple correlation analysis. We interpret the trained glass-box models directly and do not explain how they describe the importance of regressors such as LIME, SHAP and other XAI methods. In EBM, importance is considered as the average of the absolute predictive value of each variable for the training dataset. Therefore, the importance (IMP) for EBM is calculated as follows:

$$\text{IMP}_{\text{EBM}}(i) = \frac{1}{P} \sum_{p \in P} |f_i(x_i^{(p)})| \tag{6}$$

while $x_i^{(p)} \in \mathbf{X}^{(I \times P)}$ denotes a variable of one example out of the dataset $\mathbf{X}^{(I \times P)}$ with P data points consisting of I explanatory variables. We experimented with varying numbers of variable interactions and in our case got the best results when we advised the model not to look for pairwise interactions in the data. For NAM, the variable importance is calculated similarly but the average output of every neural network is considered as well, that is computed as:

$$\overline{f}_i(x_i) = \frac{1}{R_i} \sum_{r \in R_i} f(x_{i,r}) \tag{7}$$

where $x_{i,r}$ is a possible value of variable i and R_i is the set of all possible (categorical) values of variable i. For NAM's variable importance, the average output is subtracted from the absolute predictive value of every variable as:

$$\text{IMP}_{\text{NAM}}(i) = \frac{1}{P} \sum_{p \in P} |f_i(x_i^{(p)}) - \overline{f}_i(x_i)|. \tag{8}$$

Figure 5 shows the importance of each individual predictor across all models tested for predicting the recovery rate of corporate bonds. High values indicate high relative importance of that variable in predicting the model's recovery rate (i.e., it has a large impact on the outcome). Correspondingly, low values indicate that the respective predictor is unimportant. In particular, the fact that the bonds are zero-coupon bonds has an impact on the recovery rate. Moreover, having a convertible bond or bonds from the communications sector also factors in. Both models find these three parameters important.

Looking at the importance of individual macroeconomic variables in Figure 5, we find that in particular the unemployment rate over the past five weeks is considered important by both models. The mortgage rate and the inventory-to-sales ratio are considered more important by the EBM than the NAM. Seniority is considered by both models, but we note that the NAM only considers the senior secured attribute. EMB also considers the seniority subordinated characteristic to be important. When looking at the industry variables, we find a high importance. We see a high importance of whether the respective industry is in distress. This is consistent with the results of Acharya et al. (2007) who find a strong relationship between industry distress and creditor repayment. It is noticeable that NAM considers less variables from the dataset for the recovery rate prediction than EBM. The NAM is more selective than the EBM in the characteristics to be considered.

Figure 5: Variable importance by model for all attributes used for out-of-time prediction. The characteristics are grouped by their category indicated by the left sided color stripe. The higher the value, the higher the importance of a variable evaluated by the corresponding model. The values are scaled such that the most influential variable per model is valued to 1.0.

It is noticeable that some of the important variables found are highly correlated with the recovery rate, as Figure 4 shows. In particular, the variables *Zero coupon*, *Convertible coupon* and *Industry distress 2* are worth mentioning here. It can be seen that there are some variables that are considered by the models to be very important for predicting the recovery rate, and others that are considered less or not at all. This is already a first step in the interpretation of the machine learning models. We can immediately read from the models, without any further tools, which variables are taken into account and to what extent. However, the question arises how these variables influence the result. We will discuss that in the next.

5.3 Variable Influences from Interpretable Machine Learning Methods

Complex machine learning models, especially neural networks, are often black-box models whose functions cannot be accurately interpreted. This has the disadvantage of preventing financial institutions from explaining their methods for estimating corporate bond recovery rates. Kellner et al. (2022) and Bastos & Matos (2022) take a step in this direction by explaining machine learning models. However, they do so not by interpreting the models themselves, but by using XAI methods. These are methods that explain the functionality of the models externally by analyzing the model predictions given parameters. According to Rudin (2019), these methods have the drawback of often providing explanations that do not match up with the original model's computations. In the mentioned works the variable influences are explained. The variable influences show the impact of the values of a variable on the outcome and illustrate the relationship between recovery rates and a particular regressor, such as bond characteristics, seniorities or macroeconomic variables. In this way, it is possible to understand how each variable affects the model's predictive outcome. For example, one can ask the question: How does seniority or the fact that it is a zero-coupon bond affect the recovery rate? This can be done with XAI methods such as those used by Bastos & Matos (2022), but it does not show the exact learned relationship because it is generated only by observation and not by interpretation. To interpret the influence of each variable, the learned functions f_i or h_i must be visualized as shown in Section 3. This evaluation can be performed for GAM-based models such as EBMs and NAMs. In our approach, the influence graphs are taken directly from the learned relationships of the models. They are not created by XAI methods, as in Bastos & Matos (2022), but are an exact representation of the learned relationships between the variables and the recovery rate.

Figure 6 and Figure 7 show the variable influences based on GAM functions for the most important determinants of EBM and NAMs. The variables with the highest importance for both NAM and EBM combined according to Figure 5 were used. It should be noted that variables that have no significance have

no effect on prediction considering the respective model. The score values in Figures 6 and 7 show the strength of the influence of each predictor value on the prediction of recovery rate. A positive score value indicates a strong positive influence. A negative score indicates that variables associated with this value decrease the prediction of recovery rate.

In terms of seniority levels, we see a very positive influence of recovery rates for secured bonds, whereas (senior) subordinated bonds tend to have a lower recovery rate. As with all dummy variables, it is noticeable that only one of the two values has a relevant score. This can be interpreted as follows: The recovery rate prediction is obtained by summing the results of all variable inputs; that is, the results of the functions f_i or h_i. A positive score indicates a positive influence on the recovery rate for certain seniority level. For example, if a bond is secured, the recovery rate is increased with a score of 9.58 using EBM. If it is not secured, the recovery rate prediction is not increased or even decreased. The subordinated bonds have the lowest average recoveries on average. Further, we see a result as in the correlation analysis in Figure 4. This finding supports the results of Varma & Cantor (2005), Altman & Kishore (1996), Schuermann (2004), Jankowitsch et al. (2014), Nazemi & Fabozzi (2018), Gambetti et al. (2019), Bastos & Matos (2022), and Nazemi et al. (2022) indicating that secured bonds recover more than unsecured and subordinated bonds.

Acharya et al. (2007) report that when an industry is in distress, defaulted bonds from this industry have a significantly lower recovery rate. We can confirm these results by looking at the dummy variables *Industry distress 1* and *Industry distress 2*. We find that firms in an industry that is in distress receive a lower score and are therefore more likely to have a lower recovery rate, which is consistent with the findings of previous studies (see, for example Acharya et al. (2007), and Nazemi & Fabozzi (2018)). The NAM shows a strong positive influence on the recovery rate prediction, if the industry is not under distress. This influence is less pronounced with EBM. Next, we consider bond characteristics. These contain important variables for assessing the recovery rate of corporate bonds. In particular, we see a very positive impact on the recovery rate for insured and zero-coupon bonds for the EBM. This is also in line with the correlation. In contrast, we see a significant decrease in the recovery rate for convertible coupon bonds. We find the same for retail notes and variable coupon bonds.

Further, we find that utilities have a higher value for the recovery rate, which is consistent with the results of Altman & Kishore (1996), Varma & Cantor (2005), Jankowitsch et al. (2014), Nazemi et al. (2017), and Gambetti et al. (2019), while corporate bonds from consumer goods, communications, the industrial as well as the materials sector have a lower recovery rate assessment. Finally, we analyze the continuous

macroeconomic variables mortgage rate and unemployment rate (number of civilians unemployed for less than five weeks). We find that a lower mortgage rate, in particular for NAM, leads to a increase in the recovery rate prediction. However, we also realized that significantly negative association between the recovery rates and number of civilians unemployed for less than five weeks, which is consistent with Nazemi & Fabozzi (2018). An unemployment rate below this critical point results in a score of 15 for NAM and 10 for EBM resulting in a higher recovery rate prediction. This confirms the findings of Tobback et al. (2014) for corporate data who state a positive relationship between the unemployment rate and loss given default.

Comparing the results of both models, we identify similarity in most variables, while the NAM is showing stronger influences distributed over fewer variables. The interpretation of the machine learning models for recovery prediction produced the following result: both by the importance of the predictors, and even more by the consideration of the influences, the models used for recovery prediction can be interpreted very well. The interpretations provide insight into what attributes of a bond or other information related to the defaulted bond are used by the models and how a recovery rate is predicted, which is a benefit. Our approach shows accurate visualization of the relationships learned by the models without the use of XAI methods. Therefore, the relationships do not have to be approximated and represent the exact learned relationship and thus the exact mode of action in predicting the recovery rate. Through this interpretation, we were able to identify strong influencing factors on the recovery rate prediction as well as its influence (i.e. in a positive or negative direction). This analysis would not have been possible if black-box models had been used. The use of glass box models such as EBM or NAM should therefore be considered rather than the usage of complex black-box models, because glass-box models are interpretable and perform equally well as black-box models that are not interpretable. Our influence diagrams in Figures 6 and 7 are similar to the PDPs used by Bastos & Matos (2022) to explain the significance of variables in black-box models, but these are post-hoc measures, whereas our diagrams consist of learned functions and thus describe the model directly. Therefore, the diagrams show how a variable value affects the outcome recovery rate prediction. For NAM and EBM, we find very similar influences of the variables on the recovery prediction. While for NAM some variables are not taken into account very much (i.e. they show little impact on the scores), the impacts of the other variables are larger in absolute terms than for EBM. However, a common trend can be observed in both methods and is the same across the models.

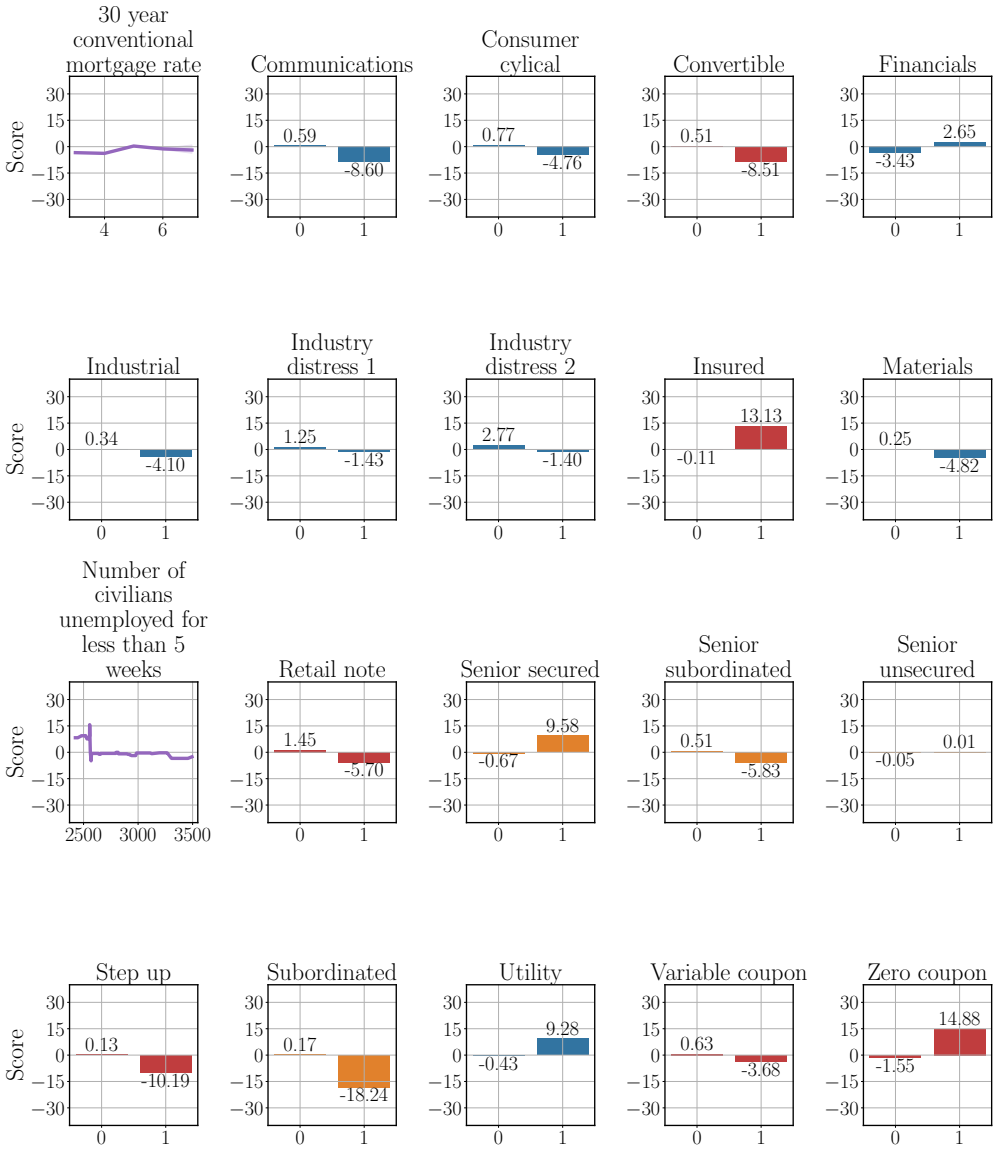

Figure 6: Variable influences of EBM for out-of-time prediction (red: bond characteristics; green: company variables; blue: industry variables; yellow: seniority variables, purple: macroeconomic variables). The panels show the sensitivity of creditor recovery rates (vertical axis) to each attribute.

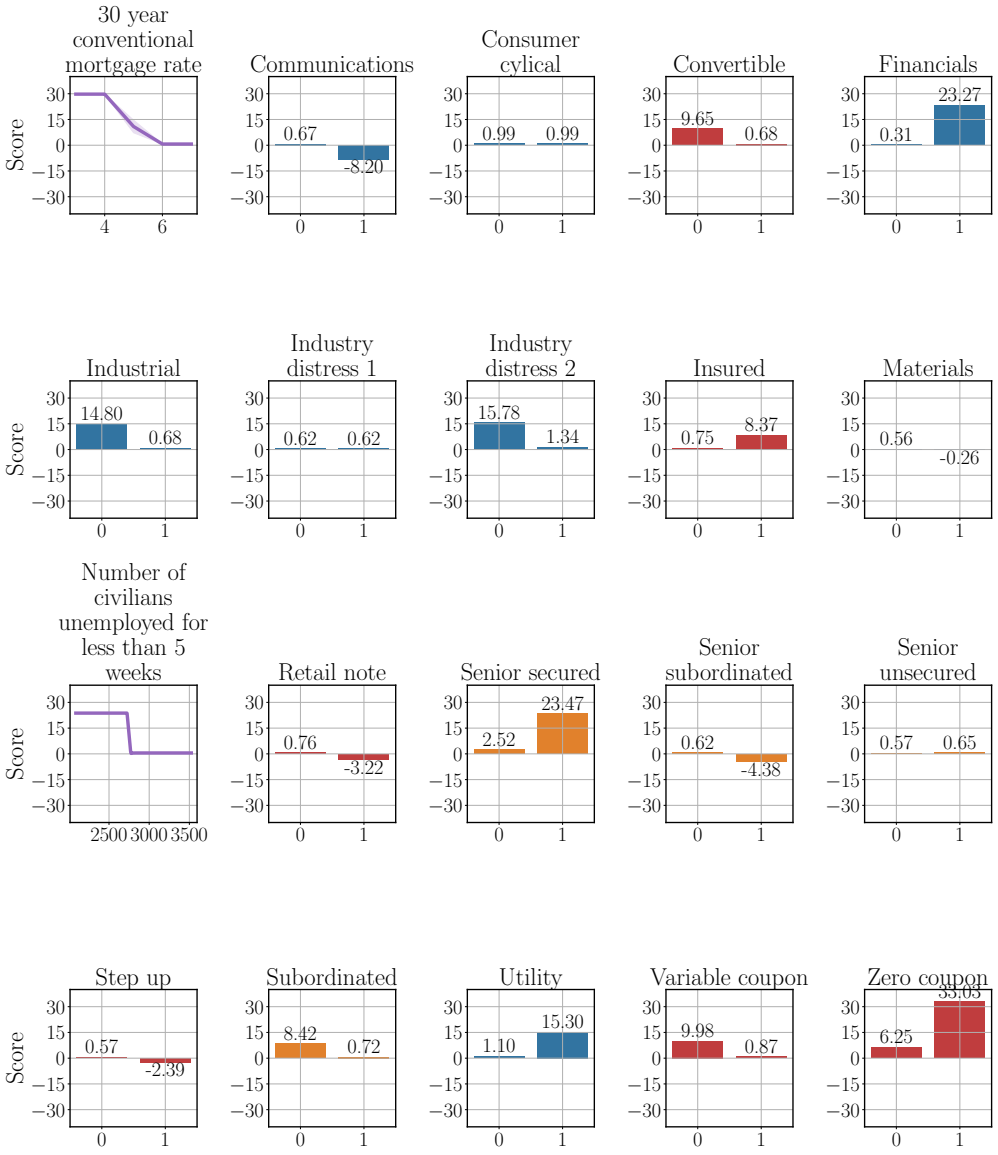

Figure 7: Variable influences of NAM for out-of-time prediction (red: bond characteristics; green: company variables; blue: industry variables; yellow: seniority variables, purple: macroeconomic variables).

6 Conclusion

Because of the complex relationships between the corporate bond recovery rates and drivers of recovery rates, recent studies have shown that machine learning methods are able to significantly outperform other linear methods to predict recovery rate. Risk analysts cannot comprehend how specific predictions have been made using machine learning approaches. On the other hand, meeting regulatory obligations requires insight into the methods financial institutions use. Moreover, to make the user or affected person more familiar with the results and functionality of a model and to increase the acceptance of machine learning models. Interpretable machine learning models are becoming increasingly important.

In this paper in contrast to the existing literature (which focuses on predictions using machine learning algorithms), we propose interpretable machine learning algorithms for predicting recovery rates of U.S. corporate bonds from 2001-2016. We compare our suggested models with various machine learning techniques to estimate recovery rates of defaulted corporate bonds over this period. Our empirical results show that it is possible to use intrinsically interpretable glass-box models for creditor recovery prediction while achieving better prediction performance. At the same time, it is possible to explain the results of these interpretable models more easily than with black-box models. We exhibit the influence of each predictors on the estimation of the recovery rate. It became clear that bond-specific variables in particular, as well as the company's industry determinants as well as macroeconomic values, have a higher influence on the result. The results of interpreting the importance and influences of individual variables show similarities to the correlation of those attributes with the recovery rate. In summary, intrinsic interpretation provides an advantage for predicting creditor recovery rates because the models can now be understood without relying on post-hoc testing procedures. As a consequence, the intelligible models can be better verified and more confidence can be gained in the models.

The results found in this paper can help to estimate recovery rates by uncovering patterns and hidden relationships between explanatory variables and recovery rates, calculating capital requirements, and managing bond portfolio credit risk. The success of interpretable machine learning methods for recovery rate estimation covers the main drawback of machine learning for practitioners and academics. Based on our study, we see promising opportunities for applying interpretable machine learning in finance and economics.

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*, 787–821.

Agarwal, R., Melnick, L., Frosst, N., Zhang, X., Lengerich, B., Caruana, R., & Hinton, G. E. (2021). Neural additive models: Interpretable machine learning with neural nets. *Advances in Neural Information Processing Systems*, *34*, 4699–4711.

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*, 2203–2228.

Altman, E. I., & Kalotay, E. A. (2014). Ultimate recovery mixtures. *Journal of Banking and 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*, 57–64.

Arık, S. , & Pfister, T. (2021). Tabnet: Attentive interpretable tabular learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*, 6679–6687.

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

Bastos, J. A., & Matos, S. M. (2022). Explainable models of credit losses. *European Journal of Operational Research*, *301*, 386–394.

Bellotti, A., Brigo, D., Gambetti, P., & Vrins, F. (2021). Forecasting recovery rates on non-performing loans with machine learning. *International Journal of Forecasting*, *37*, 428–444.

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

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

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

FICO (2018). Explainable machine learning challenge, https://community.fico.com/s/explainable-machine-learning-challenge.

Frye, J., Frye, & Jon (2000). Depressing recoveries. Federal Reserve Bank of Chicago.

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

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

Hastie, T., & Tibshirani, R. (1986). Generalized additive models. *Statistical Science*, *1*, 297–310.

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

Jacobs, M., & Karagozoglu, A. K. (2011). Modeling ultimate loss-given-default on corporate debt. *Journal of Fixed Income*, *21*, 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*, 155–177.

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

Kaposty, F., Kriebel, J., & Löderbusch, M. (2020). Predicting loss given default in leasing: A closer look at models and variable selection. *International Journal of Forecasting*, *36*, 248–266.

Kellner, R., Nagl, M., & Rösch, D. (2022). Opening the black box – quantile neural networks for loss given default prediction. *Journal of Banking and Finance*, *134*.

Lou, Y., Caruana, R., Gehrke, J., & Hooker, G. (2013). Accurate intelligible models with pairwise interactions. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *Part F1288*, 623–631.

Martins, A. F. T., Com, A. M., & Astudillo, R. F. (2016). From softmax to sparsemax: A sparse model of attention and multi-label classification.

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

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences of the United States of America*, *116*, 22071–22080.

Nazemi, A., Baumann, F., & Fabozzi, F. J. (2022). Intertemporal defaulted bond recoveries prediction via machine learning. *European Journal of Operational Research*, *297*, 1162–1177.

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

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*, 664–675.

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

Nori, H., Jenkins, S., Koch, P., & Caruana, R. (2019). Interpretml: A unified framework for machine learning interpretability. *CoRR*, *abs/1909.09223*.

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

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*, 206–215.

Schuermann, T. (2004). Financial institutions center what do we know about loss given default? *Wharton Financial Institutions Center Working Paper*, *No 04-01*.

Sopitpongstorn, N., Silvapulle, P., Gao, J., & Fenech, J. P. (2021). Local logit regression for loan recovery rate. *Journal of Banking and Finance*, *126*, 106093.

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

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

Vaughan, J., Sudjianto, A., Brahimi, E., Chen, J., & Nair, V. N. (2018). Explainable neural networks based on additive index models. *The RMA Journal*, (pp. 40–49).

Wang, H., Forbes, C. S., Fenech, J.-P., & Vaz, J. (2020). The determinants of bank loan recovery rates in good times and bad-new evidence. *Journal of Economic Behavior and Organization*, *177*, 875–897.

Yang, Z., Zhang, A., & Sudjianto, A. (2021). Gami-net: An explainable neural network based on generalized additive models with structured interactions. *Pattern Recognition*, *120*, 108–192.

Yao, X., Crook, J., & Andreeva, G. (2015). Support vector regression for loss given default modelling. *European Journal of Operational Research*, *240*, 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*, 679–689.

Online Appendices

Appendix A TabNet Masks

TabNet by Arık & Pfister (2021) constructs masks that are arrays denoting the importance of each feature of each decision step. The mask of each step i is an array of the form $\mathbf{M}[\mathbf{i}] \in \mathbb{R}^{(B \times I)}$ with batch size B and number of features I. Masking is done by multiplying the feature vector $\mathbf{X}^{(B \times I)}$ with the mask. Masks of each step i are obtained from the previous step using the attentive transformer and sparsemax proposed by Martins et al. (2016), an activation function that is similar to the traditional softmax, but outputs sparse probabilities. Sparsemax has the following form:

$$\text{sparsemax}(\mathbf{z}) = \begin{cases} t-1 & t > 1 \\ \frac{t-1}{2} & -1 \leq t \leq 1 \\ -1 & t < -1 \end{cases} \tag{A.1}$$

where $\mathbf{z} = (t, 0)$.

The mask of each step i is calculated from the outcome of the previous step $\mathbf{a}[\mathbf{i}-1]$ by

$$\mathbf{M}[\mathbf{i}] = \text{sparsemax}(\mathbf{P}[\mathbf{i}] \cdot f_i(\mathbf{a}[\mathbf{i}-1])) \tag{A.2}$$

where f_i is a trainable function using fully-connected layer, followed by batch normalization. $\mathbf{P}[\mathbf{i}]$ is the prior scale term, that is calculated by

$$\mathbf{P}[\mathbf{i}] = \prod_{j=1}^{i} (\gamma - \mathbf{P}[j]). \tag{A.3}$$

with γ being a relaxation parameter.

Appendix B Epririal Analysis

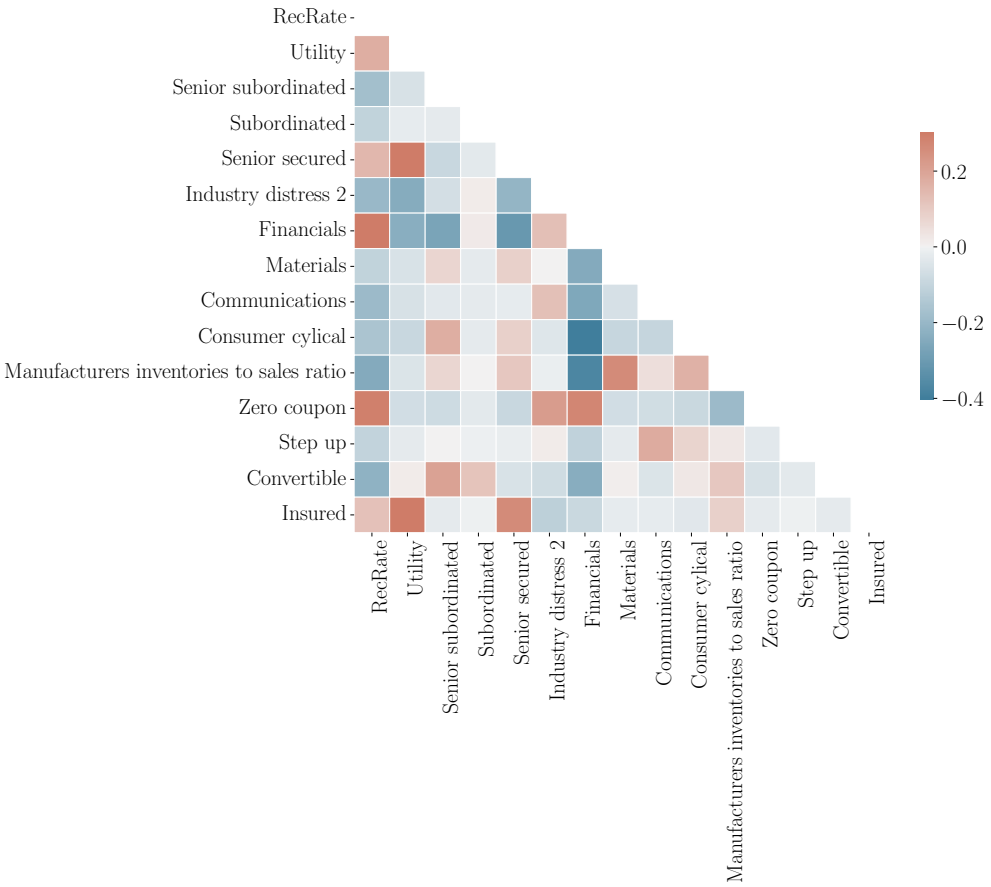

Figure A.1: Correlation matrix of selected features.