2020

Predicting Recovery Rates using Machine Learning Algorithms: the Relative Usefulness of Alternative Methods

yiping lu

Claremont McKenna College

Predicting Recovery Rates using Machine Learning Algorithms: the Relative Usefulness of Alternative Methods

submitted to
Professor George Batta
and
Professor Weiqing Gu

by
Yiping Lu

for
Senior Thesis
Spring 2020
May 11, 2020

Contents

1 Abstract

This study evaluates the performance of linear model trees to forecast recovery rates of defaulted bonds. The linear model trees are built based on regression trees, with a linear regression model in each leaf. I use bond characteristics, firm characteristics, industry indicators, and macroeconomic indicators as explanatory variables. The relevance of explanatory variables is assessed using the Mutual Information Feature Selection method. The results show that the linear model trees present better out-of-sample forecasts of recovery rates in comparison with some other widely-used models.

2 Introduction

The financial crisis has highlighted the importance of better approaches in credit risk modeling. The Basel Committee on Banking Supervision issued the Basel Accords, a series of banking supervision accords that ensure minimum capital requirements for banks. In Basel 1, the minimum required capital was calculated based on the risky assets owned by the bank and was used to minimize the credit risk. Some criticisms over Basel 1 include the limited differentiation of credit risk and the static measure of default risk. The Basel 2 Accord was published in 2004 to improve the regulations, where it allows banks to use their internal approaches to calculate regulatory capital. Basel 2 introduces three risk factors in the internal models to estimate credit risk: the probability of default, the loss given default, and the exposure at default, so loss given default has become a more important measure. The loss given default is defined as a percentage of exposure that a financial institution loses when a bond defaults, and recovery rate equals one minus the loss given default.

This paper focuses on the modeling of recovery rates using machine learning methods. Traditionally, the linear regression model has been applied to predict the recovery rates (see, e.g. Altman and Kishore (1996); Altman et al. (2005); Acharya et al. (2007)). A distinct approach is to use non-parametric models, where no assumption about distribution is being made. The flexibility of many non-parametric models, such as regression trees and support vector machines, outperforms parametric regressions. A more complex model developed based on regression trees, the linear model tree, is developed by Quinlan (1992) where linear regression models are used in the leaves of decision trees. Thus, linear model trees are expected to outperform regression trees. In the area of credit risk, Siami, Gholamian, Basiri, Fathian (2011) find that an application of a locally linear model tree increases predictive accuracy in the credit scoring area. In this paper, I will investigate the performance of linear model trees to forecast recovery rates.

The remainder of this paper is organized as follows. Section 2 includes the literature review on recovery rates. An overview of the data set of recovery rates and explanatory variables is provided in Section 3. Section 4 reviews the linear regression, the support vector regression,

the decision tree, the random forest, and the linear model tree techniques. Section 5 presents the main results of these models. Section 6 contains the conclusion.

3 Literature Review

In the past, credit risk modeling has been dominated by the probability of default, where the recovery rate is assumed to be independent of the default rate. Frye (2000) modeled recovery rates based on the assumption that recovery rates are dependent on a systematic factor, the state of the economy, and they further discovered a strong negative correlation between default rates and recovery rates using Moody's Default Risk Service database for the 1982-1997 period. The negative correlation and the dependence of recovery rates on the state of the economy are further studied and affirmed by the findings of Jarrow (2001) and Acharya, Bharath, and Srinivasan (2003). Altman, Brady, Resti, and Sironi (2005) find that the performance of the economy is less predictive than some studies suggest. They conclude that recovery rates are a function of supply and demand for bonds, with default rates playing an important role.

Altman and Kishore (1996) analyze the strong effects of seniority and industry affiliation on recovery rates. They find that bonds issued by companies in public utilities and chemical has the highest average recoveries, using a data set of over 700 defaulted bonds for the 1978-1995 period. Hanson and Schuermann (2004) reinforces the evidence for the impact of seniority and industry affiliation, where higher recovery rates come from the utility sector and technology and telecommunication firms. Acharya, Bharath, and Srinivasan (2007) provide evidence for the impact of industry-wide distress on to recovery rates at default. R. Jankowitsch (2014) provides an analysis of the determinants of recovery rates, where they use a very comprehensive list of instrument- and firm-specific variables to analyze a broader set of default events.

In addition to the determinants of recovery rates, modeling of recovery rate prediction has developed during the last several decades. A straightforward approach is to apply linear regression (Altman and Kishore (1996); Altman et al. (2005); Acharya et al. (2007)). Methods using beta regression models are suggested by Calabrese and Zenga (2010) and X. Huang and Oosterlee (2011). Bastos (2010) suggests that the results of a non-parametric predictive model, the regression tree, outperform a parametric model, the fractional response regression. Loterman et al. (2012) provides evidence that non-linear models, such as support vector machines and artificial neural networks, give more accurate results than traditional linear approaches. Yao, Crook, and Andreeva (2015) reaffirms the outstanding predictive power of support vector methods compared to linear regression and fractional response regression.

4 Data

To compare the predictive accuracy of models, I use a comprehensive list of variables that are significant when predicting recovery rates based on literature. Most of my data is from Moody's Default & Recovery Database (DRD). While Moody's DRD provides data for 500,000+ debts and 50,000+ global corporate and sovereign entities, I use data on debts trading within the US market. The sample consists of 3,728 bonds that defaulted between 1972 and 2019.

I collect firm characteristics from Compustat and CRSP, including balance sheet and income statement items and trading information. In order to predict the recovery rate at the time of the default event, I collect data for fiscal years prior to the default. I am only able to collect information for about 70% of firms, because some companies are subsidiaries of their parent companies and financial reports of their parent companies do not reflect the performance of themselves. The data set collected from CRSP is also used to calculate industry characteristics, such as industry return and volatility.

The fourth data set I use is obtained from the Federal Reserve Economic Data (FRED), which provides important macroeconomic information such as treasury rates, corporate yield spreads, unemployment rates, and gross domestic product.

4.1 Recovery Rate

In Moody's DRD database, there are three types of recovery rates provided: the discount liquidity, the discount settlement, and the trading price. I use the recommended discounted recovery price, which is the rate recommended by Moody's Investor Service based on internal research standards. The price is determined based on either the liquidity, settlement, or trading price method, discounted back from a trading date to the last date cash paid, using the defaulted instrument's effective interest rate.

The price is expressed as a proportion of the par value of a bond, which is scaled from 0 to 1. The distribution of the recovery prices in this study follows a bimodal distribution, which implies that most of the defaulted bonds have recovery close to full repayment, or there is no recovery at all.

Figure 1 shows the probability density of recovery rates, where the y-axis represents the probability density function for the kernel density estimation and the total area under the curve integrates to one.

4.2 Bond Characteristics

The data set in my study contains 3,728 observations.

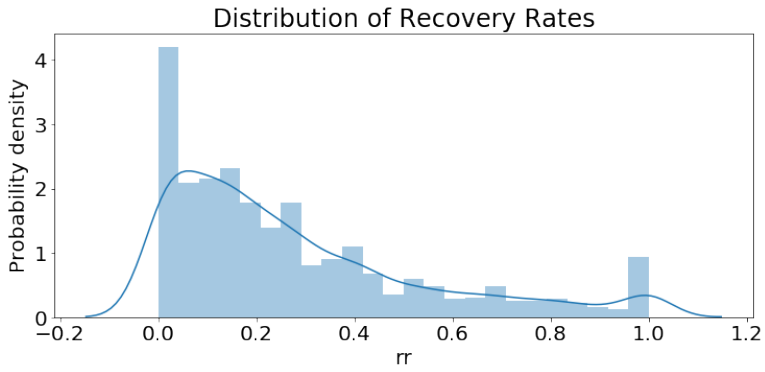

Figure 1: Recovery Rates Distribution

4.2.1 Debt class

There are 11 types of debts represented in the data set: bank credit facility (BCF), convertible/exchange/debenture (CON), enhanced equipment trust (EET), equipment trust (EQT), first mortgage bond (FMB), industrial revenue bond (IRB), pass-through certificate (PAS), preferred stock (PRF), regular bonds (REG), surplus notes (SHF), secured lease obligation bond (SLB). The majority of debts are regular bonds and bank credit facility, which represent 35% and 52% of total debts. A box-plot of the different debt types found in the database is shown in Figure 2. The box-plot shows the minimum, the first and third quantile, the maximum and the median of each type.

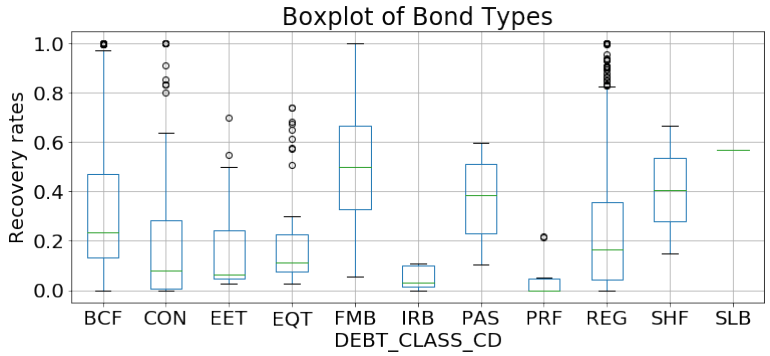

Figure 2: Debt class box-plot

4.2.2 Debt seniorities

There are 6 types of debt seniorities represented in the data set: revenue bonds (IRB), multiple seniorities, junior subordinated (JS), preferred stock (PS), subordinated (SB), senior subordinated (SR), senior unsecured (SU), and senior secured (SS). The majority of seniorities are senior secured and senior unsecured, which represent 46% and 32% of total debts. A box-plot of the different seniorities found in the database is shown in Figure 3. The box-plot shows the minimum, the first and third quantile, the maximum and the median of each seniority.

Seniority refers to the order of repayment in the event of a sale or bankruptcy of the issuer. Based on the box-plot, senior secured debt holders have higher recovery rates compared to other types of debt holders.

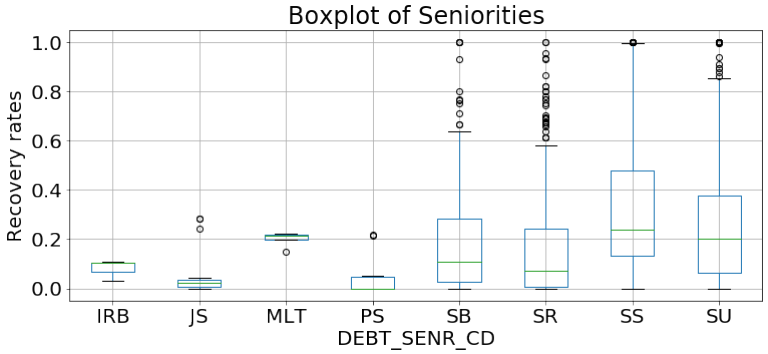

Figure 3: Seniority box-plot

4.2.3 Industries

There are 11 industries classified by Moody's 11 Code in the data set: capital industries, consumer industries, energy and environment, media and publishing, nonbank finance, real estate investment trust, retail and distribution, sovereign and public finance, technology, transportation, utilities, and unassigned. The majority of companies operate in capital industries, which represent 28% of total companies. A box-plot of the different industries found in the database is shown in Figure 4. The box-plot shows the minimum, the first and third quantile, the maximum and the median of each industry.

According to the box-plot in Figure 4, the utilities industry has the highest recovery rate while real estate investment trust has the lowest recovery rate.

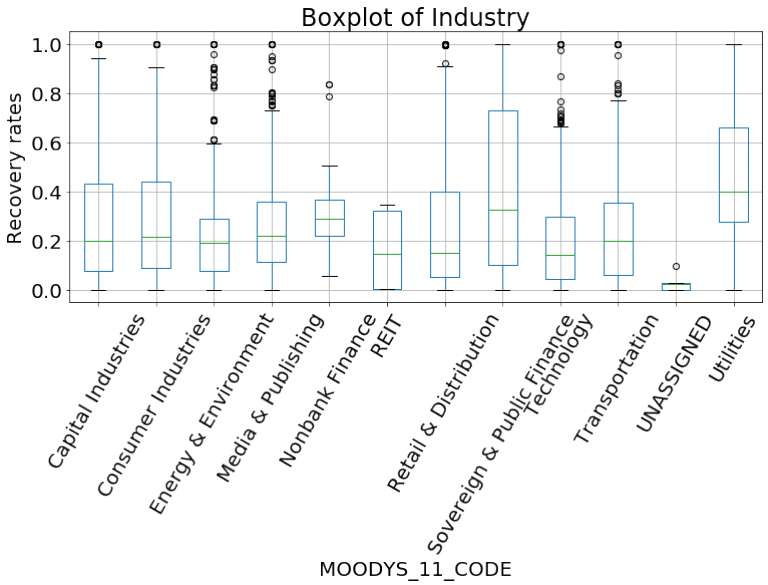

Figure 4: Industry box-plot

4.2.4 Collateral rankings

Moody gives 7 rankings to collateral, where instruments in each event are ranked in relation to each other based on the structure prior to default, taking into consideration collateral and instrument type. For example, instruments with intellectual property as collaterals tend to have a ranking of 3, while most instruments with all current assets as collaterals have a ranking of 1. The majority of rankings are in 1 and 2, which represent 43 and 39% of total debts. A box-plot of the different industries found in the database is shown in Figure 5. The box-plot shows the minimum, the first and third quantile, the maximum and the median of each collateral ranking.

According to the box-plot in Figure 5, it is clear that instruments with first ranking have the highest recovery rate while seventh ranking instruments has the lowest recovery rate.

4.2.5 Outcomes

There are 3 possible outcomes of each default event, where companies are liquidated, emerged, or acquired. The majority of companies are emerged, which represent 67% of total event outcomes. A box-plot of the different outcomes found in the database is shown in Figure 6. The box-plot shows the minimum, the first and third quantile, the maximum and the median of each outcome.

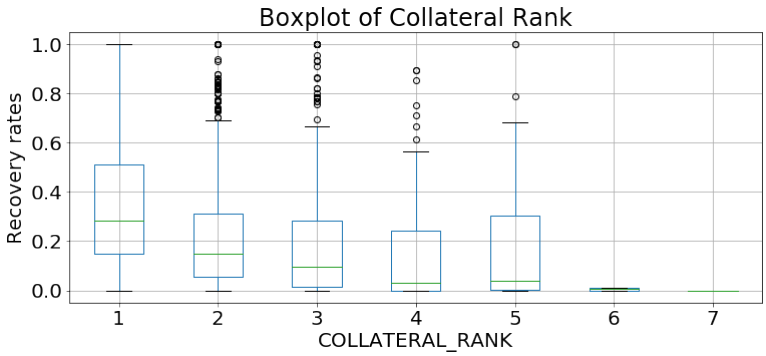

Figure 5: Collateral Ranking box-plot

The resolution-type variable provides a more specific description of how the issuer came out of default, which is classified into 14 categories. The majority of resolution types are reorganization plan confirmed and emerged from Chapter 11, which represent 35% and 23% of total types. A box-plot of the different resolution types found in the database is shown in Figure 6. The box-plot shows the minimum, the first and third quantile, the maximum and the median of each resolution type.

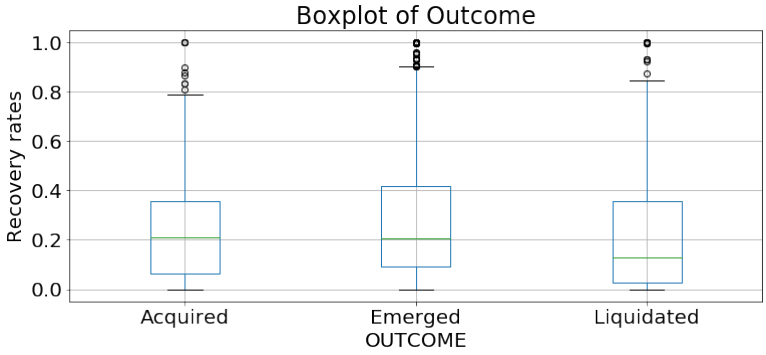

Figure 6: Outcome box-plot

4.2.6 Ratings

There are 18 types of credit ratings given by Moody to each debt one year prior to the default. The majority of companies are in B2&B3 and Caa1&Caa2, which represent 31% and 25% of total debts. A box-plot of the different ratings found in the database is shown in Figure 7. The box-plot shows the minimum, the first and third quantile, the maximum and the median of each rating.

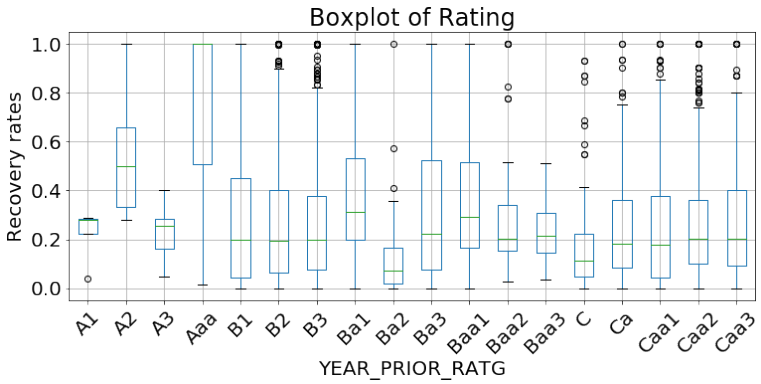

Figure 7: Ratings box-plot

4.3 Firm Characteristics

Firm characteristics are accounting measures obtained from Compustat and CRSP, for the fiscal years prior to the default event.

According to Acharya et al. (2007), I use the logarithmic transformation of total assets to represent the firm size. Other accounting ratios are based on Jankowitsch et al. (2014):

$$\text{Default barrier} = \frac{\text{Short-term debt} + 1/2\text{Long-term debt}}{\text{Total assets}} \tag{1}$$

$$\text{LTD issuance} = \frac{\text{Long-term debt}}{\text{Total debt}} \tag{2}$$

$$\text{Profitability} = \frac{\text{EBITDA}}{\text{Total assets}} \tag{3}$$

$$\text{Intangibility} = \frac{\text{Intangible assets}}{\text{Total assets}} \tag{4}$$

$$\text{Receivables} = \frac{\text{Total receivables}}{\text{Total assets}} \tag{5}$$

4.4 Industry Indicators

Chuang et al. (2019) considered two equity market variables, industry return, and industry volatility. I retrieve stock prices between 1972 to 2019 from CRSP and group by industry using the SIC 2-digit code. I calculated the monthly average return and 12-month return volatility for each industry. The two industry indicators are added to each stock for the month the default event happened. The underlying assumption is that industry-wide events reflected in the equity market have an impact on default and recovery of individual firms.

Name	Mean	Std.Dev	Min	Median	Max
Industry return (%)	0.01	0.10	-0.50	0.01	0.57
Industry volatility (%)	0.04	0.10	0.00	0.03	1.27
Federal Funds rate (%)	3.37	2.75	0.07	2.50	14.78
Slope (%)	1.68	1.27	-1.25	1.85	3.85
Aaa Corporate yield spread (%)	6.57	1.89	2.94	6.55	15.27
Baa Corporate yield spread (%)	7.68	1.89	3.77	7.87	17.18
Unemployment (%)	5.88	1.61	3.30	5.50	10.60
Real GDP (log)	14.78	0.39	13.47	14.81	15.50
CPI (log)	5.18	0.21	4.43	5.18	5.55

Table 1: Summary statistics for industry and macroeconomic indicators

4.5 Macroeconomic Indicators

I collect macroeconomic indicators from FRED. Jankowitsch et al. (2014) defined the slope of the yield curve as the difference between the Federal Funds rate and the ten-year US Treasury yield. They argued that the Federal Funds rate and the slope of the term structure are indicators of the state of the business cycle. They also considered the Federal Funds rate as the relevant short-term interest rate to avoid default risk and illiquidity. I also employ other variables such as the logarithmic transformation of real GDP, unemployment rate, and CPI, which are commonly used in previous literature, including Altman et al. (2005), Gambettia et al. (2019), and Chuang et al. (2019).

A complete list of predictors can be seen in Appendix.

5 Methodology

5.1 Prediction Using Linear Regression Model

5.1.1 Data preparation

Figure 8 provides number and percentage of empty values, following a descending order: In all the models, I replace numeric empty values with mean values and discrete empty

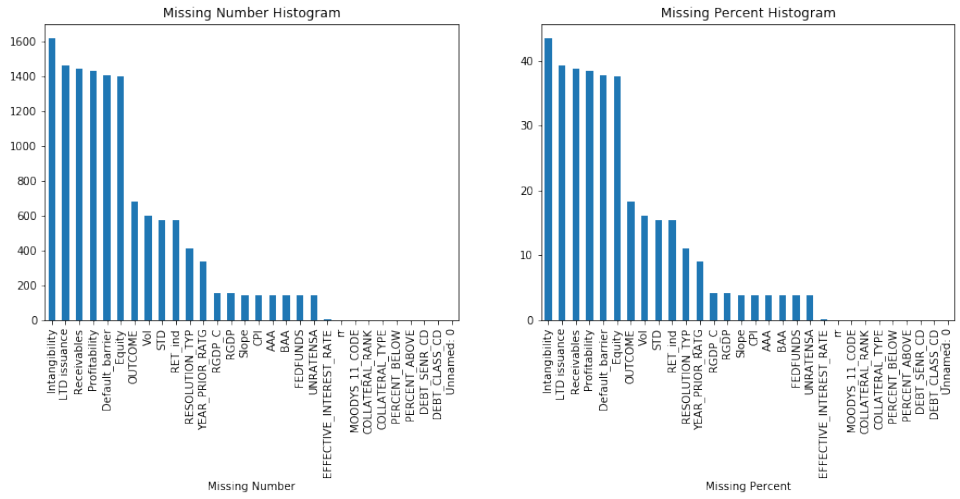

Figure 8: Missing values

values with a new variable representing missing value. I randomly split the data into the training set and the test set, where the training set contains 70% of the entire data set and the test set contains 30% of the data set.

Many machine learning models require a transformation of categorical variables to some numerical representations. SciKit-learn provides two commonly used methods: LabelEncoder and OneHotEncoder.

LabelEncoder encodes categorical variables with value between 0 and $n-1$, where n is the number of unique values in each feature, in alphabetical order. However, many features, such as industry names and debt types, do not have an order or rank. When I convert them to different numbers in a column, the model will misunderstand the data and captures the relationship based on the order of numbers. As a result, the model may imply some

correlations between the categorical features and the recovery rates, but it may not work in the test data set or other data.

OneHotEncoder is one of the machine learning methods to avoid this problem. OneHotEncoder converts each category value into a new column and assigns a 0 or 1 value to the column, where 0 denotes false and 1 denotes true.

OneHotEncoder eliminates the order issue in LabelEncoder. However, more columns are added to the data set, especially if a certain categorical variable contains too many different values.

The resulting data set of LabelEncoder consists of 28 independent variables and 1 dependent variable. The resulting data set of OneHotEncoder consists of 103 independent variables and 1 dependent variable.

5.1.2 Linear Regression

Linear regression is one of the most basic and commonly used models in financial and economic studies. Many studies on recovery rates prediction include linear regression.

The recovery rate y of a bond i is defined by the following equation:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \epsilon \tag{6}$$

where $x_1, x_2, \ldots, x_n$ are the independent variables, $\beta_0, \beta_1, \ldots, \beta_n$ are coefficients assigned to each independent variable through error minimization, and ϵ is the error term.

5.2 Support Vector Regression

Support vector regression (SVR) allows non-linear fitting through non-linear kernel functions. I use Radial Basis Function (RBF) as the kernel function to map the data to a higher dimensional feature space. SVR is based on the following function:

$$y = \sum_{i=1}^{N} (a_i - a_i^*) \cdot K(x_i, x) + b \tag{7}$$

where the RBF kernel function $K(x_i, x)$ is given by:

$$K(x_i, x_j) = exp(-\frac{||x_i - x_j||^2}{2\sigma^2}) \tag{8}$$

5.3 Decision Tree

A. Bellotti and Crook (2012) find that a regression tree outperforms the linear regression when predicting recovery rates. Regression trees have become very popular and effective

following the work of Breiman (1984). A decision tree is built by splitting the observations, which form its roots, into homogeneous subsets, based on a specific feature at each node. For categorical features, a split is done on the belonging or not to a particular class. For continuous features, a split is done based on thresholds. Each split is done by minimizing the impurity, or the sum of squared errors, which is given by:

$$Q = \sum_{i=1}^{N} (y_i - \bar{y_i})^2 \tag{9}$$

where

$$y = \frac{1}{N} \sum_{i=1}^{N} (y_i) \tag{10}$$

After a split based on feature k and threshold c, the impurity is given by:

$$Q(k, c) = \sum_{i:X_{ik} \leqslant c} (y_i - \bar{y_{k,c,l}})^2 + \sum_{i:X_{ik} > c} (y_i - \bar{y_{k,c,r}})^2 \tag{11}$$

where "l" and "r" denote "left" and "right" and $\bar{y_{k,c,l}}$ and $\bar{y_{k,c,r}}$ denote the average outcomes in the two sub-samples. The error is reduced at each split. When the splitting process finishes, the prediction for a leaf is the sample average value within each leaf of the tree. Regression trees are easy to interpret and can handle non-linearity well.

However, regression trees are often prone to over-fitting. Therefore, I use 10-fold cross-validation to evaluate the results. A K-fold cross-validation partitions training data into K equally sized sub-samples. In each fold, the algorithm uses the other K-1 sub-samples as training data and the last sub-sample as validation.

5.4 Random Forest

Random forests are an ensemble learning method that constructs a multitude of decision trees through bootstrap aggregating or bagging. The algorithm resamples the original data set into new random subsets with replacement. Then an individual regression tree is built and optimized on each of the random subsets. The final predictions of a random forest is the average of all the predictions given by each tree. Because the sub-samples are random, the trees are uncorrelated, the ensemble predictions are more accurate than any of the individual predictions. Therefore, random forests are expected to give better predictions than decision trees.

5.5 Linear Model Trees

Linear model trees combine linear regression models and decision trees to create an integrated model that may give better predictions and insights. In a typical regression tree,

the leaf nodes contain a sample mean of the training set values. Quinlan (1992) developed an M5 model tree where linear regression models are used in the leaves of regression trees. A model tree is a combination of piece-wise linear models, where the algorithm breaks the input space and assigns a linear model suitable for a sub-area. The combined model is locally accurate and handles non-linearity. It can also directly learn from continuous and discrete variables.

I built a linear model tree using linear regression functions (Equation 12) in the leaves.

$$\hat{y}_i = \beta_0 + \beta_1 \hat{y}_{i1} + \beta_2 \hat{y}_{i2} + \ldots + \beta_n \hat{y}_{in} + \epsilon \tag{12}$$

where $\hat{y}_i$ is the predicted value of recovery rate at node i.

Model	R^2	MAE	MSE	RMSE
LR-LabelEncoder	0.3732	0.2675	0.1113	0.3336
LR-OneHotEncoder	0.4406	0.246	0.0993	0.3151
SVR-OneHotEncoder	0.5602	0.1993	0.0781	0.2794
DT-OneHotEncoder	0.5879	0.1193	0.0732	0.2705
DT-LabelEncoder	0.6235	0.1179	0.0668	0.2585
RF-LabelEncoder	0.781	0.1181	0.0389	0.1972
MT-LR	0.88	0.0798	0.0213	0.146

Table 2: Results of the models

6 Analysis & Results

6.1 Evaluation Metrics

R-squared is a widely used indicator of the goodness of fit of a model. In linear regression, R can be interpreted as the proportion of variation of the dependent variable y that is explained by the independent variables $x_1, x_2, \ldots, x_n$ of the model. The higher the R-squared value, the better the model.

$$R^2 = 1 - \frac{\text{Residual Sum of Squares}}{\text{Total Sum of Squares}} \tag{13}$$

I also use the following indicators to evaluate the model:

$$\text{MSE} = (\frac{1}{n}) \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 \tag{14}$$

$$\text{MAE} = \frac{\sum_{i=1}^{n} |\hat{y}_i - y_i|}{n} \tag{15}$$

$$\text{RMSE} = \sqrt{(\frac{1}{n}) \sum_{i=1}^{n} (\hat{y}_i - y_i)^2} \tag{16}$$

6.2 Results

Table 2 provides indicators evaluate the performance of the regression prediction models:

6.2.1 LabelEncoder and OneHotEncoder

Although LR-LabelEncoder model gives a much higher in-sample R-squared than LR-OneHotEncoder (see Table 3), the out-of-sample R-squared of LR-LabelEncoder model is

Model	In-sample R^2	Out-of-sanple R^2
LR-OneHotEncoder	0.452	0.4406
LR-LabelEncoder	0.803	0.3732

Table 3: In-sample and out-of-sample results

much lower. This implies that LR-LabelEncoder model has the problem of over-fitting, which can be avoided when using OneHotEncoder. However, in tree-based models, LabelEncoder gives better results than OneHotEncoder, as suggested by the evaluation metrics of the decision tree model in Table 2. The decision tree can directly learn from categorical values by splitting the data based on belonging or not to a certain value. OneHotEncoder can lead to problems of high cardinality and sparsity. OneHotEncoder erases important categorical structure by splitting a single feature into many separate ones, so continuous variables are assigned with higher feature importance. Thus, OneHotEncoder can degrade the prediction of tree-based models.

6.2.2 Linear Regression

In linear regression model, an R-squared score of 0.4406 suggests that approximately 44.06% of the variation in the dependent variable, the recovery rate, is explained by the variables we included in the linear regression model. These indicators suggest that linear model does not give a very accurate prediction of the recovery rate we set out to predict. This result matches the previous studies, where researchers commonly obtained low R-squared when running linear regressions.

6.2.3 Support Vector Regression

From Table 2 it is clear to see that SVR outperforms LR, where approximately 56.02% of variation is explained by the model. The SVR-OneHotEncoder model improved the out-of-sample R^2 by about 27%, while reducing the errors by 19%, 21%, and 11%, respectively, compared to LR-OneHotEncode model. This may imply that there are some non-linearity in the data that cannot be explained by linear models. The limitation in my model is that I am using the simple RBF SVR, while previous studies show that improved SVR models can derive better predictions.

6.2.4 Decision Tree

Based on Table 2, the DT model with OneHotEncoder presents similar predictive accuracy with SVR model with OneHotEncoder, and DT with LabelEncoder gives better predictions than SVR and LR. DT-LabelEncoder further improves the predictions, where R^2 increases by 11% compared to the SVR-OneHotEncoder model. Further reductions in errors of 42%, 14%, and 7% are observed. Similar to the conclusion from SVR model, this may result

0.66990263	0.44240045	0.44831631	0.60217355	0.48265977
0.55423751	0.65135245	0.30777824	0.38251953	0.5464426

Table 4: 10-fold cross validation results

from the non-linearity in the data that is supported by DT and SVR. In addition, DT can directly handle categorical independent variables, where LR requires algorithms to encode the variables.

However, DT is prone to over-fitting, as suggested by the 10-fold cross-validation results below, with a mean value of 0.51 and a standard deviation of 0.11: Table 4 gives the explained variance of each fold. To eliminate the problem of over-fitting, an RF is built to ensemble the trees.

6.2.5 Random Forest

RF gives the best predictions, where approximately 78.1% of variance is explained by the model. I determine the number of trees in the forest by optimizing the explained variance of the RF model. Figure 9 shows the change of explained variance with different number of trees: Expected variance increases from 0.59, with a standard deviation of 0.07, to 0.75,

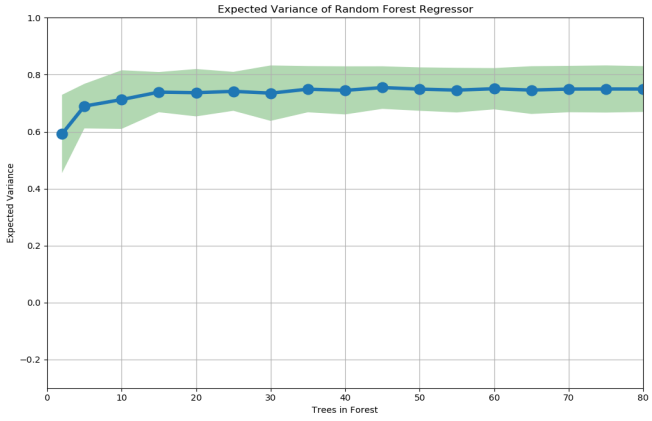

Figure 9: Expected Variance vs Number of Trees

with a standard deviation of 0.04, when the number of estimators is 35 and remains at the same level thereafter. Therefore, to improve model efficiency, I set the number of estimators to be 35. The final model has an out-of-sample R-squared value of 0.781, which outperforms the LR-OneHotEncoder model by about 77

Model	R^2	MAE	MSE	RMSE
DT-LE (max. depth= inf)	0.6235	0.1152	0.0668	0.2585
DT-LE (max. depth=4)	0.4188	0.2361	0.1032	0.3212
MT-LR	0.88	0.0798	0.0213	0.146

Table 5: Results of DT and MT

6.2.6 Linear Model Trees

As shown in Table 2, the results of the MT-LR is better than a single linear regression model and tree-based models under all the evaluation metrics. The MT-LR model improved the out-of-sample R^2 by about 13%, while reducing the MAE by 32% and MSE by 49%. While the MT-LR model gives better predictions for recovery rates, one limitation is that the algorithm is very computational-intensive, so in reality it is more expensive to implement. Due to the low efficiency in running the algorithm, I am not able to optimize the parameters in the MT-LR model. In DT and RF models, the maximum depth of a single tree is set to be None, which means the nodes are expanded until all leaved are pure or until all leaves contain less samples than the minimum number of samples required to slit an internal node. However, in MT models, I am only able to set the maximum depth to be 4 due to inefficiency. When I built a DT-LabelEncoder model with a maximum depth of 4, the results are much less accurate, as shown in Table 5:

From Table 5, we can see that the maximum depth plays an important role in model accuracy.

Thus, it suggests that predictions of MT-LR model can be further improved by optimizing parameters. However, given that the MT-LR is highly time-consuming even with a maximum depth of 4, optimization is supposed to be very expensive. There always exists the trade-off between accuracy and efficiency.

6.3 Feature Importance

6.3.1 Mutual Information

Mutual information (MI) of two random variables measures the mutual dependence between the two variables, which quantifies the amount of information obtained about one random variable though the other variable. MI is given by:

$$I(X;Y) = \int_X \int_Y p(x,y) \log \frac{p(x,y)}{p(x)p(y)} dxdy \tag{17}$$

where $p(x,y)$ denotes the joint probability density function of X and Y and $p(x)$ and $p(y)$ denote the marginal density functions. The mutual information estimates the dependence

of the joint distribution $p(x, y)$ relative to the products of the factored marginal distributions. If X and Y are independent, then $p(x, y)$ would equal $p(x)p(y)$, and $I(X; Y)$ would be zero.

In feature selection, the model maximizes the mutual information between the subset of selected features Xs and the target variable y such that:

$$S = argmax_S I(X_S; y), |S| = k \tag{18}$$

where k denotes the desired number of features.

6.3.2 Variables

The five variables with the highest MI are:

1. Percent below: this gives the percentage of the debt that is contractually subordinate to the current instrument. Debt below percentage is derived by taking the principal debt below and dividing it by the total issuer debt. Recovery rates increase as the percent below increases, which corresponding to the fact that senior debtors have more chances to recover their investments.

2. Percent above: similar to percent below, this gives the percentage of the debt that is contractually senior to the current instrument. Percent of debt above is derived by taking the principal above and dividing it by the total issuer debt. Recovery rates increase as the percent below increases.

3. Effective interest rate: this is the sum of the interest rate index, taken at last date of cash paid, and the spread over the base rate index or the fixed rate of the instrument. Recovery rates are lower when the effective interest rate is higher.

4. Standard deviation each day: this is the daily standard deviation of industry stock returns. Recovery rates decrease when standard deviation each day increases.

5. Industry 12-month volatility: this is the rolling volatility of industry stock returns over the last 12 months. Recovery rates decrease when rolling industry 12-month volatility increases.

7 Conclusion

This study evaluates the performances of five types of machine learning models and shows that the linear model tree provides the best predictions of recovery rates. The linear model tree uses linear regression functions at the leaves of a decision tree. The MT-LR with a maximum depth of 4 explains 88% of the total variation in recovery rates, using bond characteristics, firm characteristics, industry indicators, and macroeconomic indicators as explanatory variables. It is found that the percentages of debt subordinate or superior to the current instrument are particularly important. Other important variables include the effective interest rate, the standard deviation and rolling volatility of industry stock returns, and macroeconomic variables such as the CPI, the slope of the yield curve, the Federal Funds rates, and the corporate yield spreads.

By comparing the results different methods of encoding categorical features, this study provides insights into the implementation of machine learning algorithms that do not handle categorical values directly. In linear regression and support vector regression, models using OneHotEncoder gives better results than models using LabelEncoder because models will misunderstand the order of data generated by LabelEncoder. In tree-based models, LabelEncoder outperforms OneHotEncoder due to the problem of high cardinality and sparsity.

One potential drawback of linear model trees is that it is inefficient to implement. While the MT-LR model improves the accuracy of predictions, the use of model trees can have a significant computational cost, especially on large data sets. The trade-off between model accuracy and efficiency requires further study in predicting recovery rates.

8 References

Acharya, V. V., Bharath, S. T., & Srinivasan, A. (2003). Understanding the recovery rates on defaulted securities.

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 (3), 787-821.

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

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

Basel Committee on Banking Supervision. (1983). Principles for the management of credit risk. Bank for International Settlements.

Basel Committee on Banking Supervision. (2005). An explanatory note on the basel ii irb risk weight functions. Bank for International Settlements.

Bastos, J. A. (2010). Predicting bank loan recovery rates with neural networks. Technical University of Lisbon Working Paper.

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

Breiman,L. , Friedman,J.H., Olshen,R.A. & Stone,C.J. (1984): Classification and Regression Trees, Wadsworth Int. Group, Belmont, California, USA.

Calabrese,Zenga (2010). Bank loan recovery rates: Measuring and non-parametric density estimation. Journal of Banking & Finance, Volume 34, Issue 5, 903-911.

Chuang, Hui-Ching and Chen, Jau-er (2019). Industry Distress and Default Recovery Rates: The Unconditional Quantile Regression Approach.

Frye, J., et al. (2000). Collateral damage: A source of systematic credit risk.

Gambetti, Paolo & Gauthier, Geneviève & Vrins, Frederic. (2019). Recovery Rates: Uncertainty Certainly Matters. Journal of Banking & Finance. Forthcoming.

Hanson, S.G., Schuermann, T., 2004. Estimating probabilities of default. Federal Reserve Bank of New York Staff Reports No. 190.

Huang, X., & Oosterlee, C. W. (2011). Generalized beta regression models for random loss given default. The Journal of Credit Risk.

Jankowitsch, Nagler, G. Subrahmanyam (2014). The determinants of recovery rates in the US corporate bond market. Journal of Financial Economics, Volume 114, Issue 1, 155-177.

Jarrow, R. (2001). Default parameter estimation using market prices. Financial Analysts Journal, 57 (5), 75-92.

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

Quinlan,J.R. (1993) : Combining Instance-based and Model-based Learning. Proceedings of the 10th ICML. Morgan Kaufmann.

Siami, M., Gholamian, M.R., Basiri, J., & Fathian, M. (2011). An Application of Locally Linear Model Tree Algorithm for Predictive Accuracy of Credit Scoring. MEDI.

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

9 Appendix

Bond Characteristic	Firm Characteristics	Industry Indicators	Macroeconomic Indicators
Collateral type	Equity	Industry return	Real GDP
Collateral rank	Default barrier	Standard deviation each day	Unemployment rate
Effective interest rate	LTD issuance	Industry 12-month volatility	Federal Funds rate
Origination date	Profitability		BAA corporate yield spread
Maturity	Intangibility		AAA corporate yield spread
Percent of debt above	Receivables		Slope
Percent of debt below			
Prior year rating			
Maturity			
Industry			

Table 6: Summary of Explanatory Variables

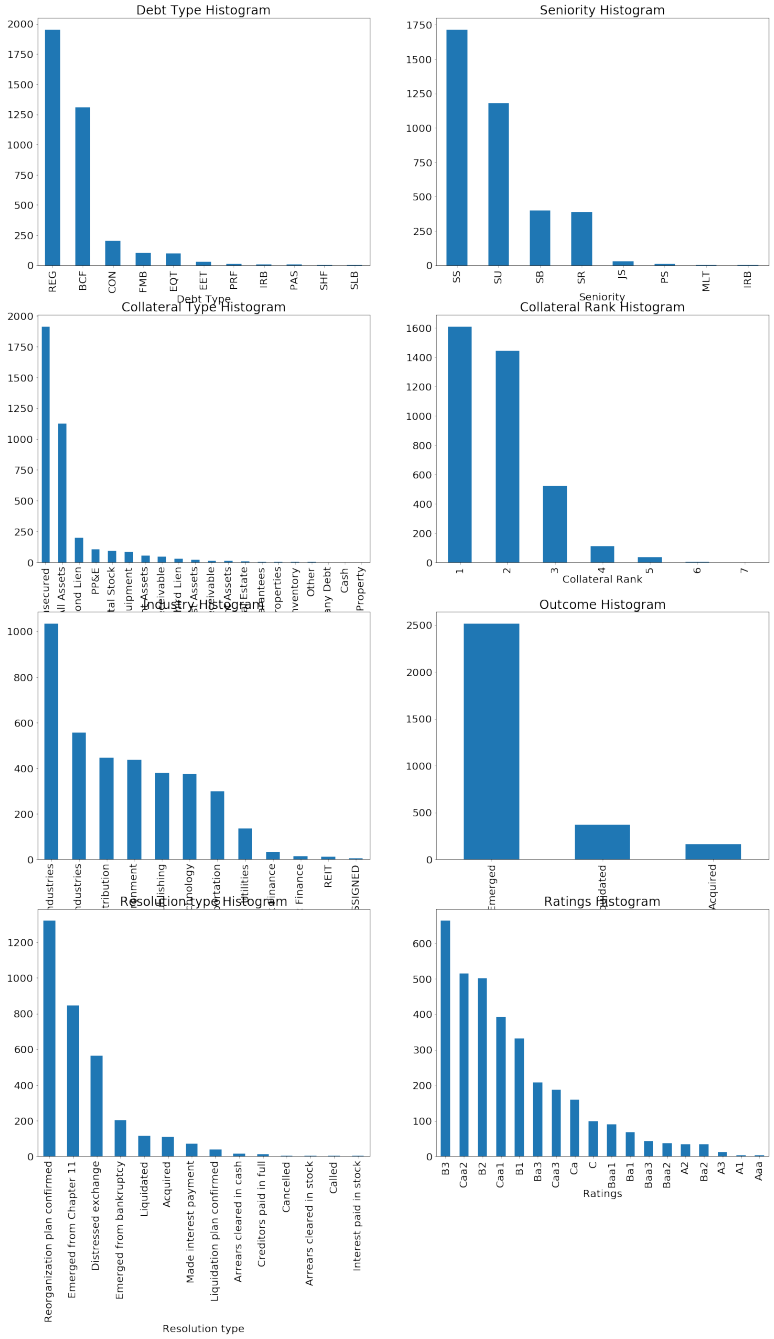

Figure 10: Categorical variables

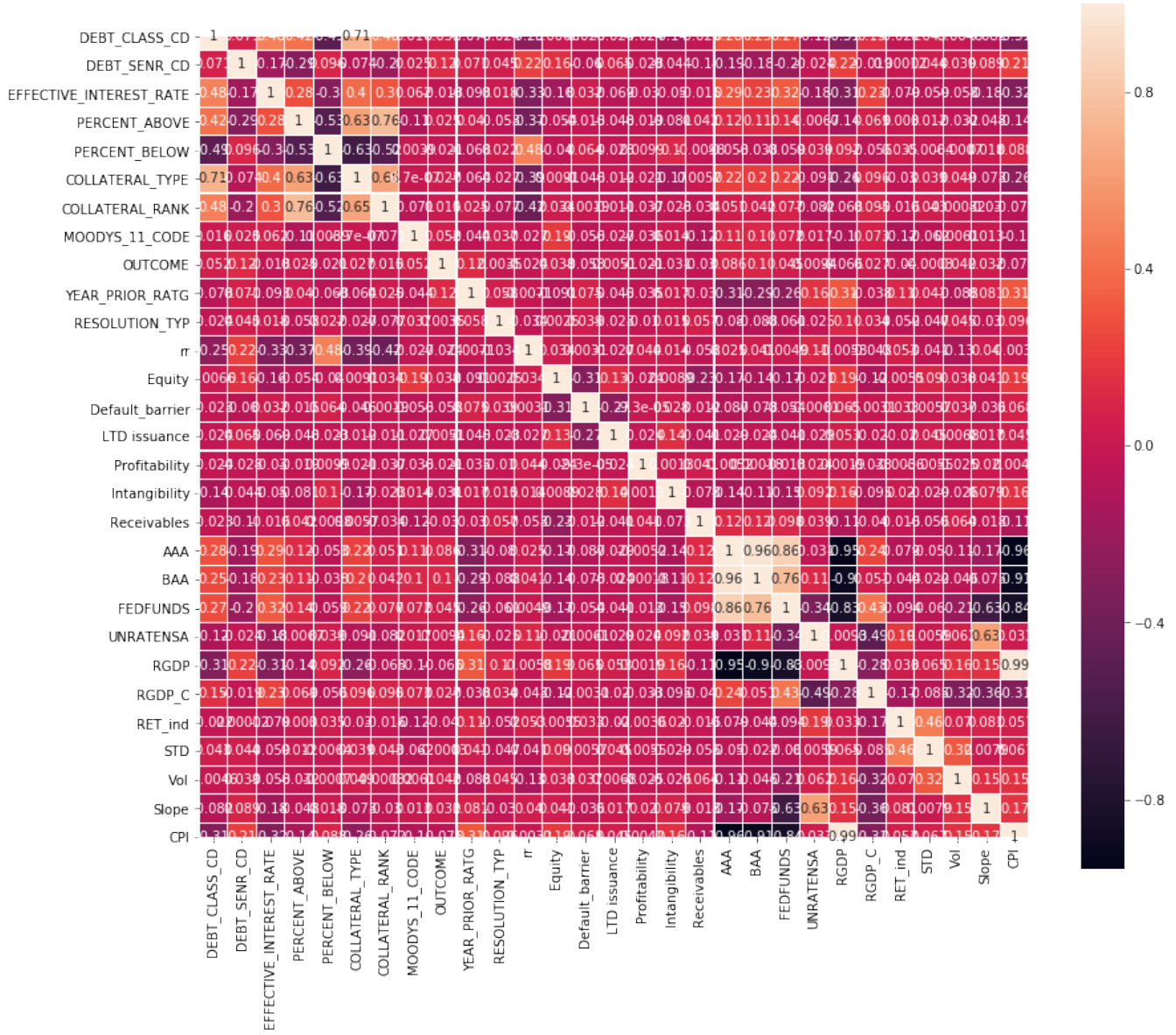

Figure 11: Correlation Map